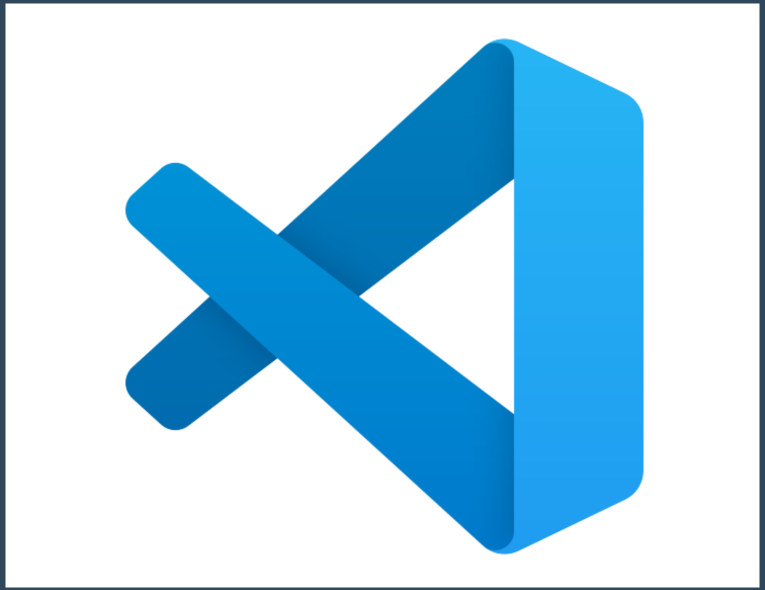
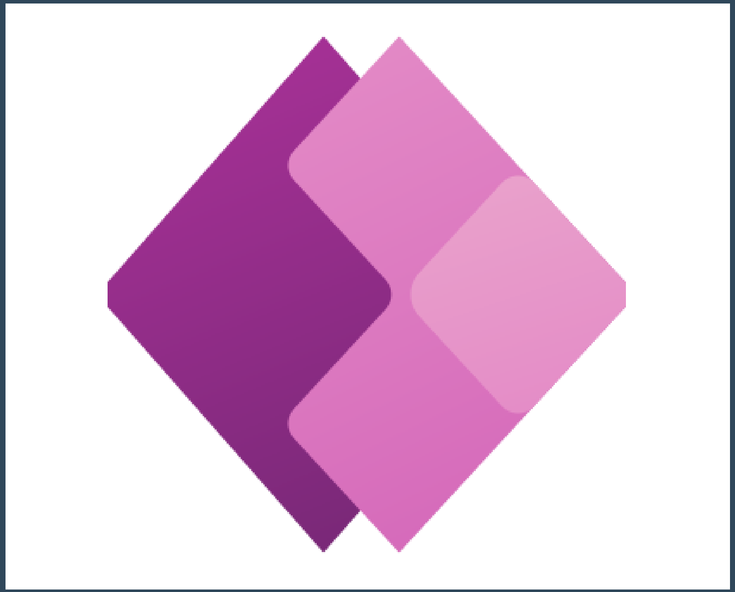
Headaches of shipping AI in products

Austin Z. Henley • Carnegie Mellon University • 9/19/2024





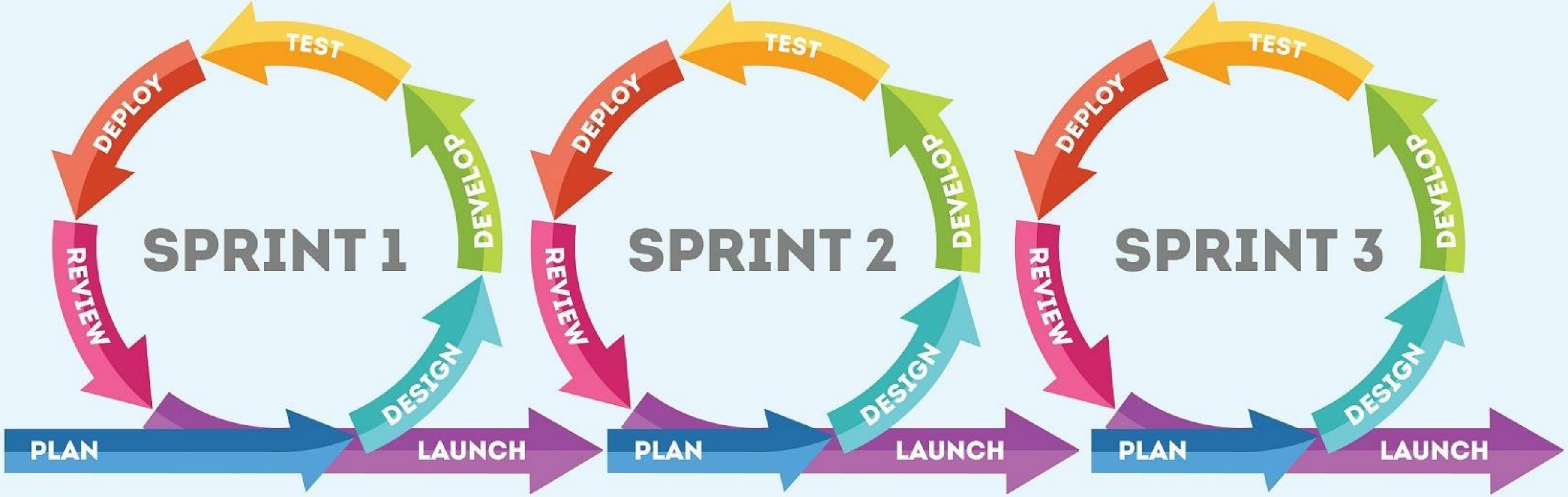
PQ

X

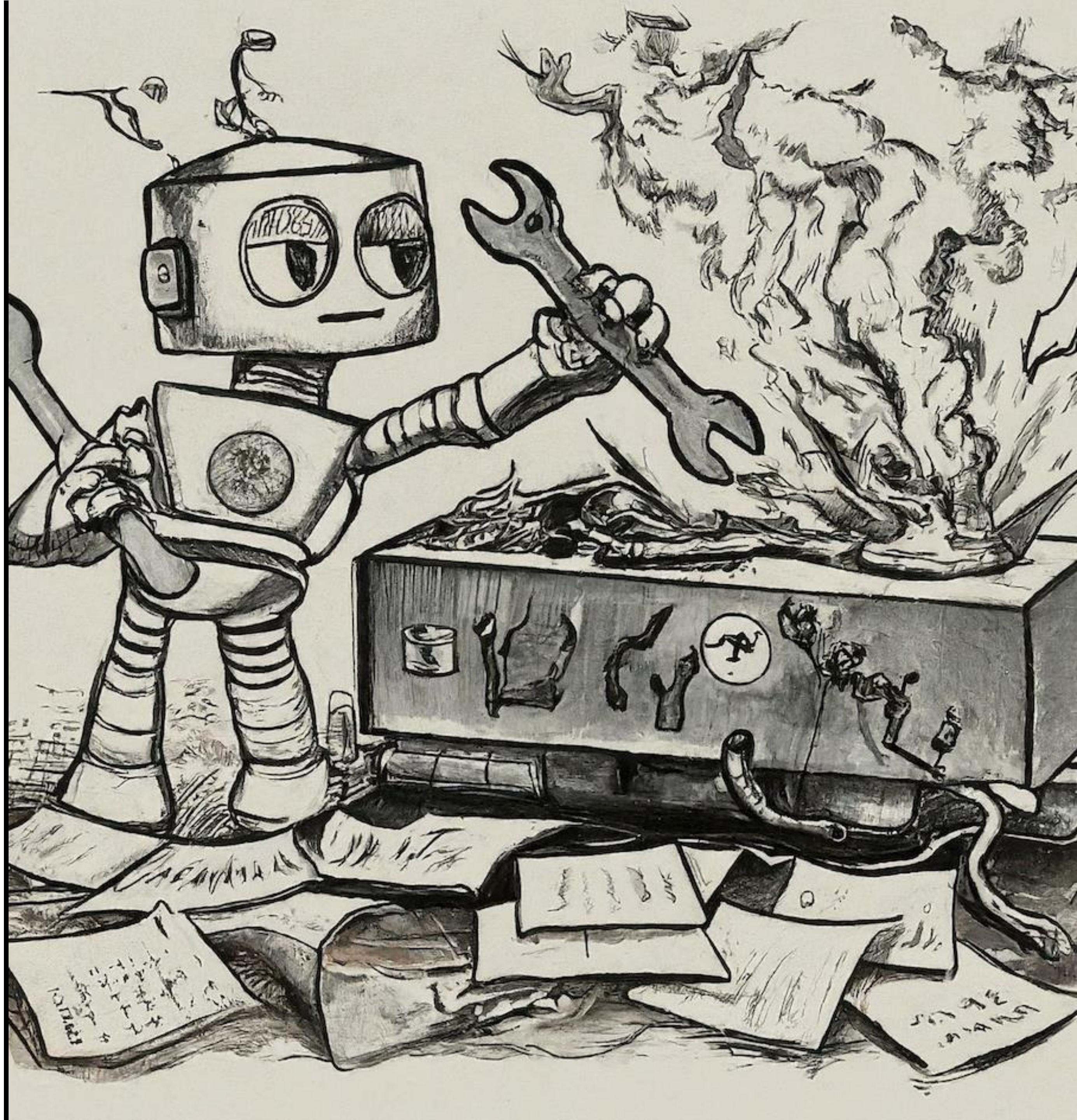
How many of you want to work on AI?

What AI features do you find useful now?

What is different for developing LLM-powered features?



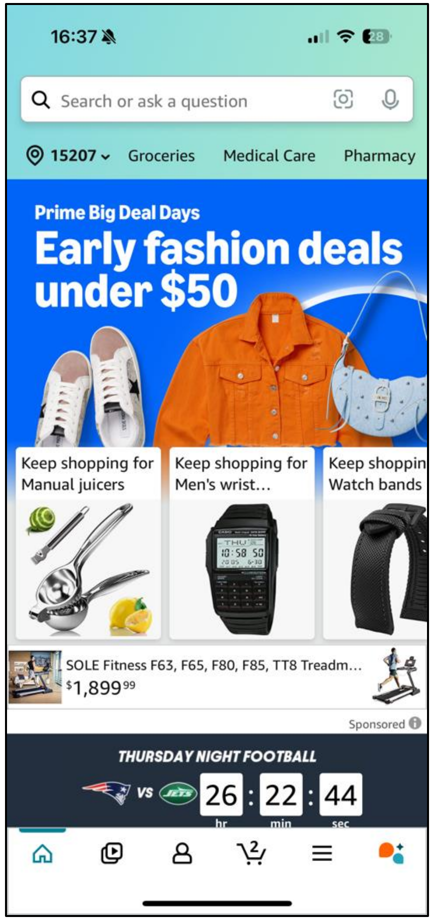
- Non-deterministic
- Costly
- Slow
- Safety
- Privacy
- Testing

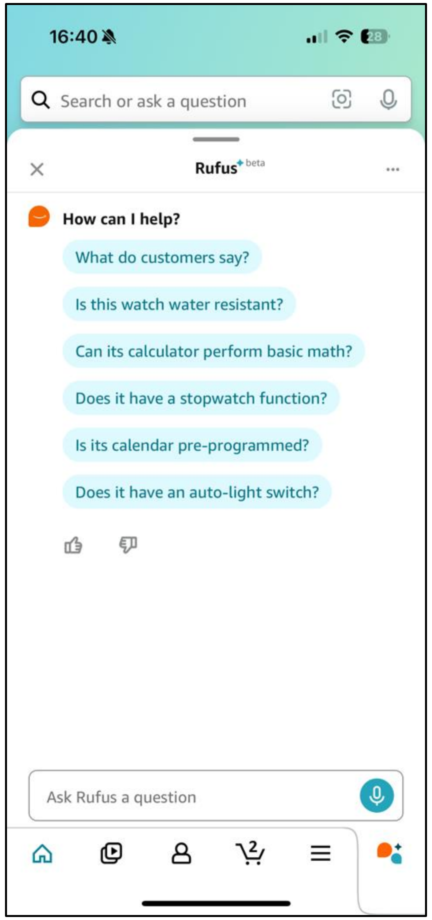


- What is the user problem we are trying to solve?
- When is the AI good enough?
- How do you know when it goes off the rails?

pRoMpT eNgInEeR

Ship it!





How can I help?

What do customers say?

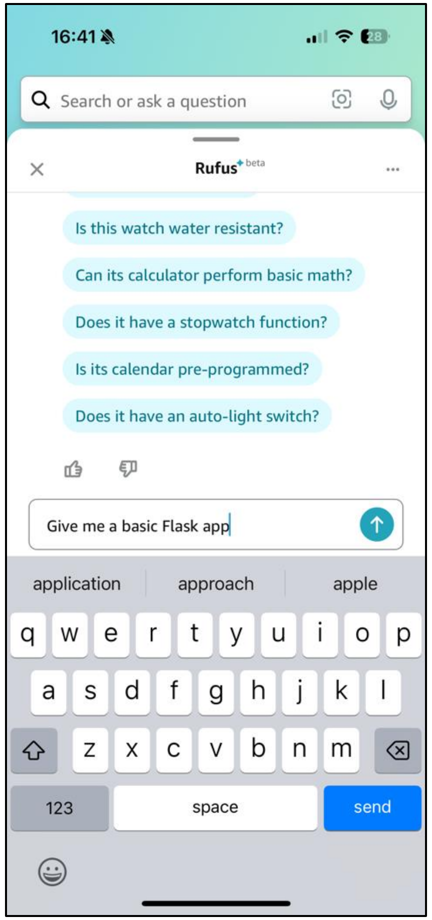
Is this watch water resistant?

Can its calculator perform basic math?

Does it have a stopwatch function?

Is its calendar pre-programmed?

Does it have an auto-light switch?



Is this watch water resistant?

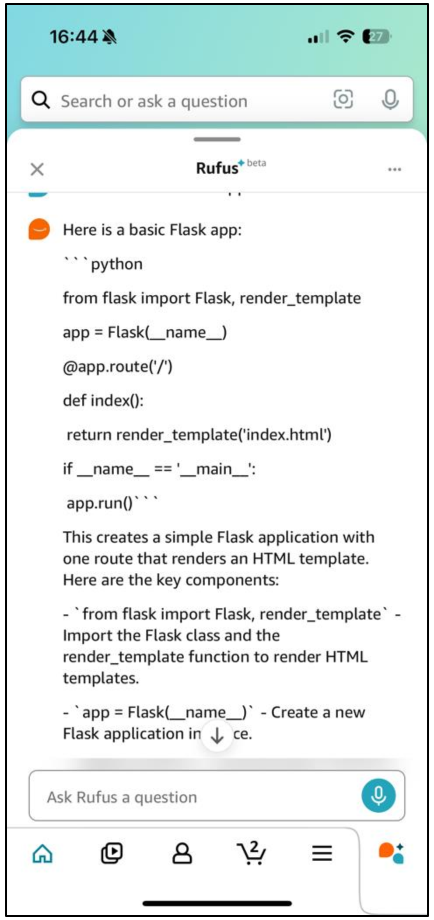
Can its calculator perform basic math?

Does it have a stopwatch function?

Is its calendar pre-programmed?

Does it have an auto-light switch?

Give me a basic Flask app



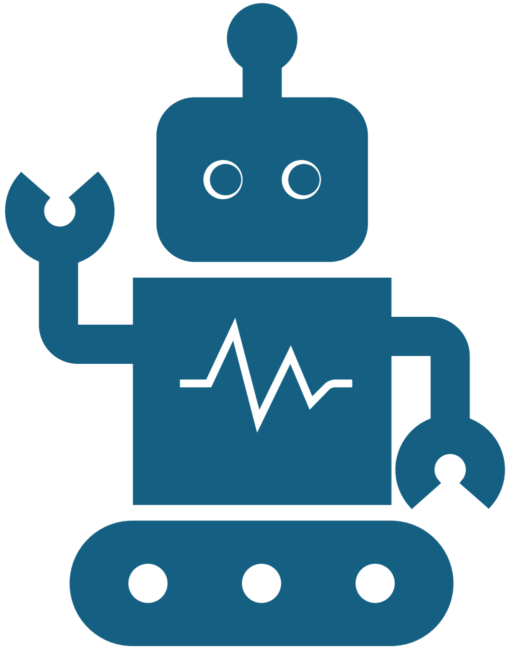
Here is a basic Flask app:

~~~
```python

from flask import Flask, render_template

app = Flask(__name__)

@app.route('/')

def index():

 return render_template('index.html')

if __name__ == '__main__':

 app.run()```
~~~

This creates a simple Flask application with one route that renders an HTML template. Here are the key components:

- `from flask import Flask, render_template` - Import the Flask class and the render_template function to render HTML templates.

- `app = Flask(__name__)` - Create a new Flask application in ce.



- When $1,000,000,000 is not enough
- We built this cool thing… can we ship it?
- Ok, we shipped… why aren’t people using it?
- Ok, we shipped… how do we know if it is helpful?
- Let’s try 50 different UIs
- How do we enable other developers to ship AI?

Before ChatGPT…

When $1,000,000,000 is not enough

Our Mission

Providing insight to the instructors: classroom analytics. 👩‍🏫

AI Coding Tutor that is indistinguishable from human instructors. 👩‍🎓

Individual student’s learning journey

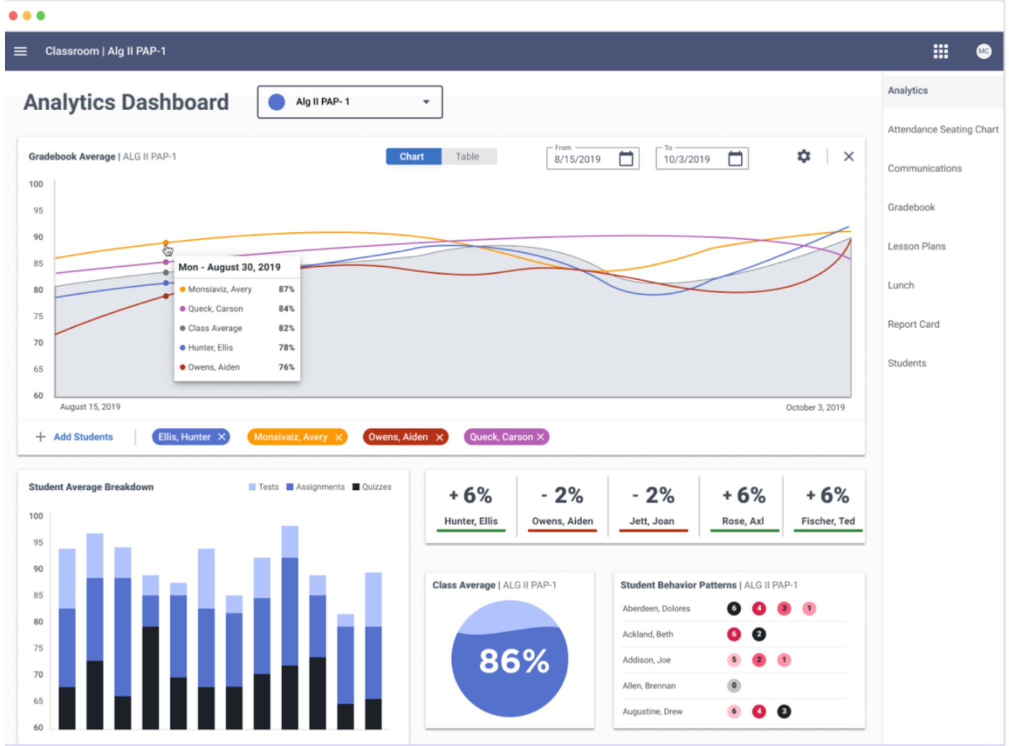
Aggregated

(class-wide) performance and pain-points

Efficiently give fair feedback

Solution: A dashboard that provides **quick summary** of students’ performance and an interface to **provide mass feedback at scale**

Goal: Assist instructors to get better insight of students’ progress

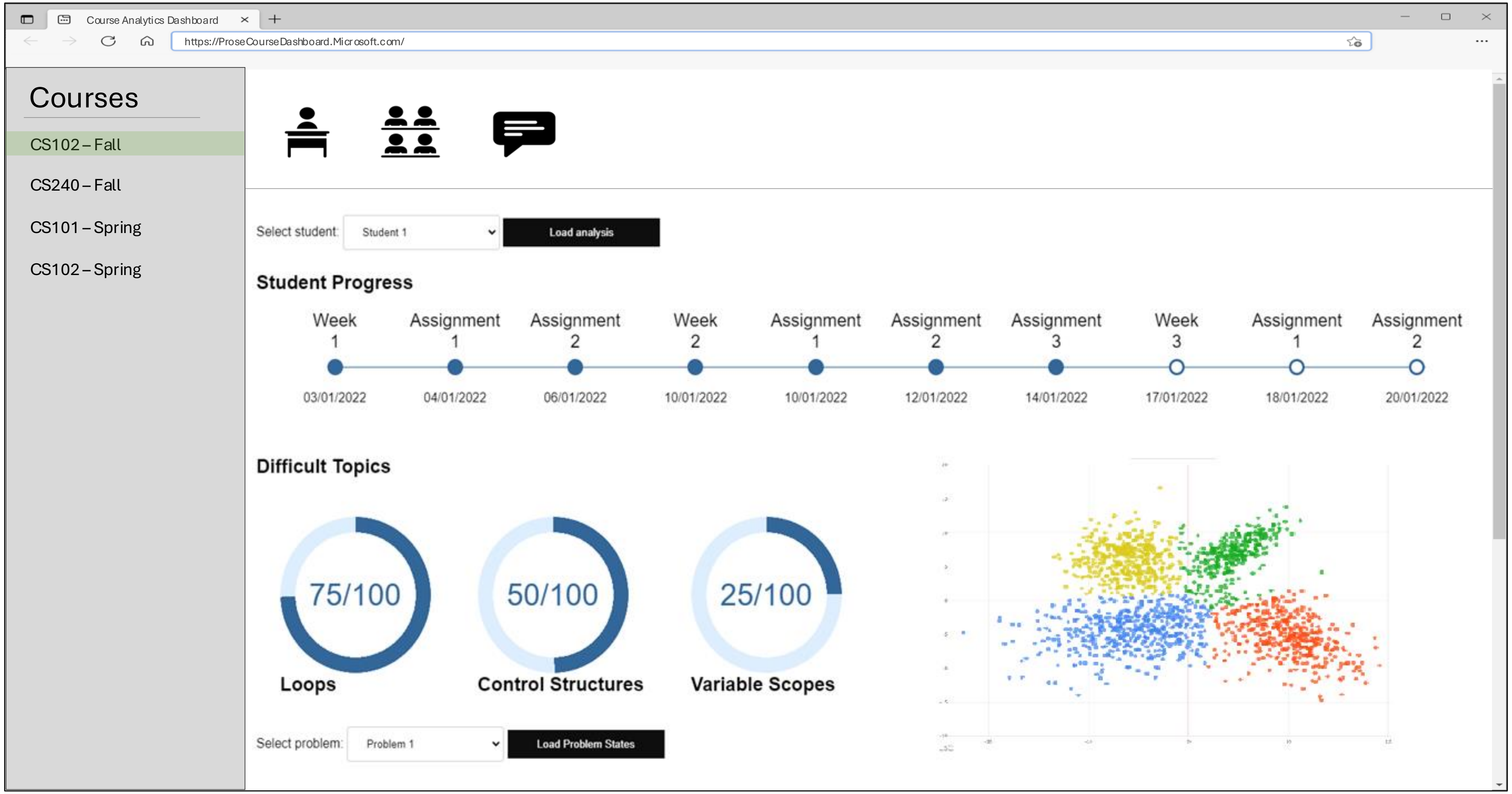


While students are our audience, teachers are our **customers**

Our tools must be built to improve the learning and teaching experience.

Source: https://dribbble.com/shots/9395179-Analytics-Dashboard-for-

Classroom analytics: Dashboard for intro to Python course



Course Analytics Dashboard
https://ProseCourseDashboard.Microsoft.com/

Courses

- CS102 – Fall
- CS240 – Fall
- CS101 – Spring
- CS102 – Spring

Select student: Student 1

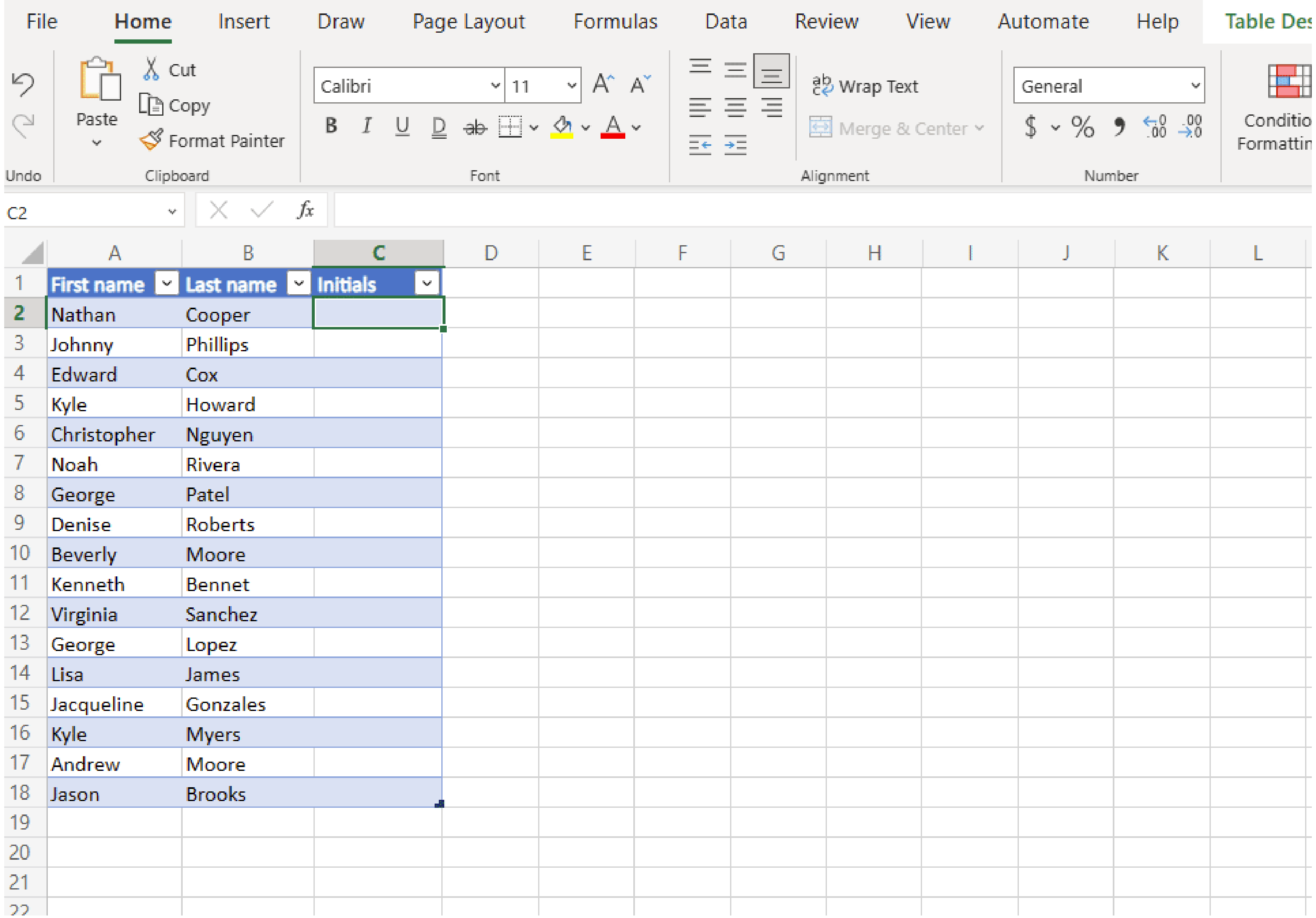
Student Progress

Week 1	Assignment 1	Assignment 2	Week 2	Assignment 1	Assignment 2	Assignment 3	Week 3	Assignment 1	Assignment 2
●	●	●	●	●	●	●	○	○	○
03/01/2022	04/01/2022	06/01/2022	10/01/2022	10/01/2022	12/01/2022	14/01/2022	17/01/2022	18/01/2022	20/01/2022

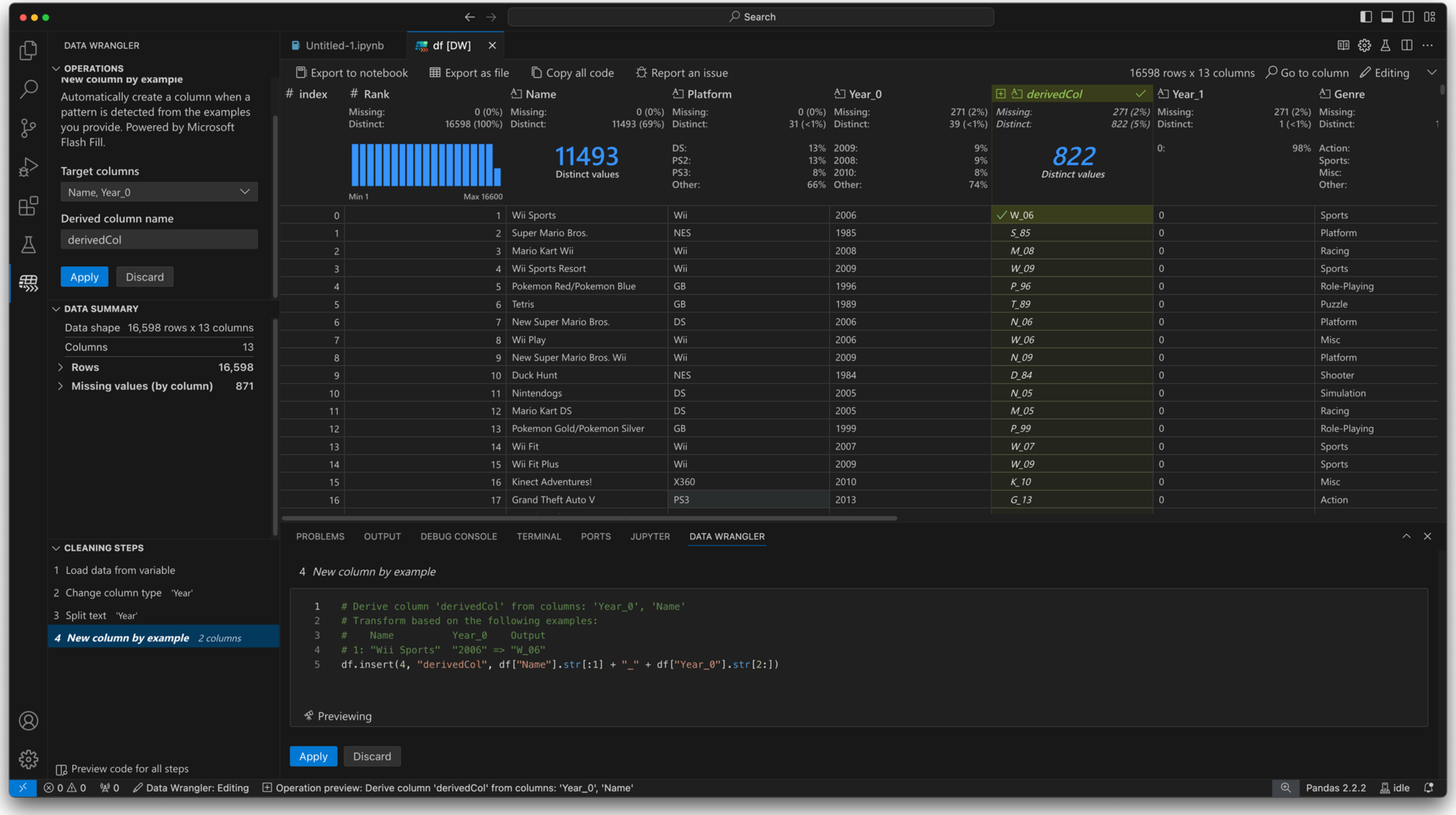
Difficult Topics

Select problem: Problem 1

We built this cool thing… can we ship it?



First name	Last name	Initials
Nathan	Cooper	
Johnny	Phillips	
Edward	Cox	
Kyle	Howard	
Christopher	Nguyen	
Noah	Rivera	
George	Patel	
Denise	Roberts	
Beverly	Moore	
Kenneth	Bennet	
Virginia	Sanchez	
George	Lopez	
Lisa	James	
Jacqueline	Gonzales	
Kyle	Myers	
Andrew	Moore	
Jason	Brooks	



OPERATIONS

New column by example

Automatically create a column when a pattern is detected from the examples you provide. Powered by Microsoft Flash Fill.

Target columns: Name, Year_0

Derived column name: derivedCol

Apply Discard

DATA SUMMARY

Data shape	16,598 rows x 13 columns
Columns	13
Rows	16,598
Missing values (by column)	871

CLEANING STEPS

1 Load data from variable

2 Change column type 'Year'

3 Split text 'Year'

4 *New column by example* *2 columns*

Preview code for all steps

16598 rows x 13 columns

index	Rank	Name	Platform	Year_0	derivedCol	Year_1	Genre
	Missing: 0 (0%) Distinct: 16598 (100%)	Missing: 0 (0%) Distinct: 11493 (69%)	Missing: 0 (0%) Distinct: 31 (<1%)	Missing: 271 (2%) Distinct: 39 (<1%)	Missing: 271 (2%) Distinct: 822 (5%)	Missing: 271 (2%) Distinct: 1 (<1%)	Missing: Distinct:
	Min 1 Max 16600	11493 Distinct values	DS: 13% PS2: 13% PS3: 8% Other: 66%	2009: 9% 2008: 9% 2010: 8% Other: 74%	822 Distinct values	0: 98%	Action: Sports: Misc: Other:
0	1	Wii Sports	Wii	2006	✓ W_06	0	Sports
1	2	Super Mario Bros.	NES	1985	S_85	0	Platform
2	3	Mario Kart Wii	Wii	2008	M_08	0	Racing
3	4	Wii Sports Resort	Wii	2009	W_09	0	Sports
4	5	Pokemon Red/Pokemon Blue	GB	1996	P_96	0	Role-Playing
5	6	Tetris	GB	1989	T_89	0	Puzzle
6	7	New Super Mario Bros.	DS	2006	N_06	0	Platform
7	8	Wii Play	Wii	2006	W_06	0	Misc
8	9	New Super Mario Bros. Wii	Wii	2009	N_09	0	Platform
9	10	Duck Hunt	NES	1984	D_84	0	Shooter
10	11	Nintendogs	DS	2005	N_05	0	Simulation
11	12	Mario Kart DS	DS	2005	M_05	0	Racing
12	13	Pokemon Gold/Pokemon Silver	GB	1999	P_99	0	Role-Playing
13	14	Wii Fit	Wii	2007	W_07	0	Sports
14	15	Wii Fit Plus	Wii	2009	W_09	0	Sports
15	16	Kinect Adventures!	X360	2010	K_10	0	Misc
16	17	Grand Theft Auto V	PS3	2013	G_13	0	Action

4 *New column by example*

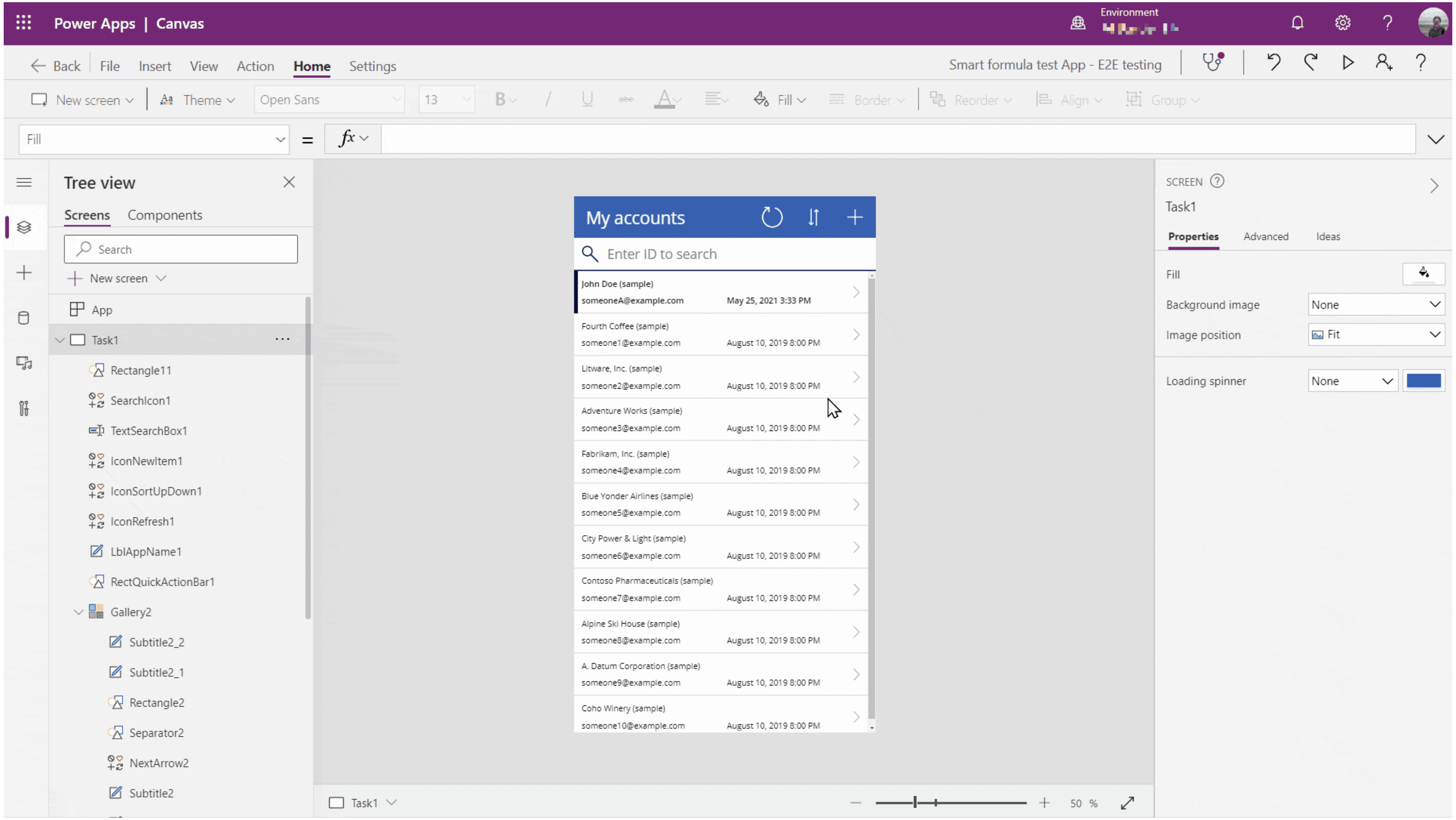
```
# Derive column 'derivedCol' from columns: 'Year_0', 'Name'
# Transform based on the following examples:
#    Name           Year_0    Output
# 1: "Wii Sports"  "2006" => "W_06"
df.insert(4, "derivedCol", df["Name"].str[:1] + "_" + df["Year_0"].str[2:])
```

Previewing

Apply Discard

Operation preview: Derive column 'derivedCol' from columns: 'Year_0', 'Name'

Ok, we shipped… why aren’t people using it?



John Doe (sample)
someoneA@example.com May 25, 2021 3:33 PM

Fourth Coffee (sample)
someone1@example.com August 10, 2019 8:00 PM

Litware, Inc. (sample)
someone2@example.com August 10, 2019 8:00 PM

Adventure Works (sample)
someone3@example.com August 10, 2019 8:00 PM

Fabrikam, Inc. (sample)
someone4@example.com August 10, 2019 8:00 PM

Blue Yonder Airlines (sample)
someone5@example.com August 10, 2019 8:00 PM

City Power & Light (sample)
someone6@example.com August 10, 2019 8:00 PM

Contoso Pharmaceuticals (sample)
someone7@example.com August 10, 2019 8:00 PM

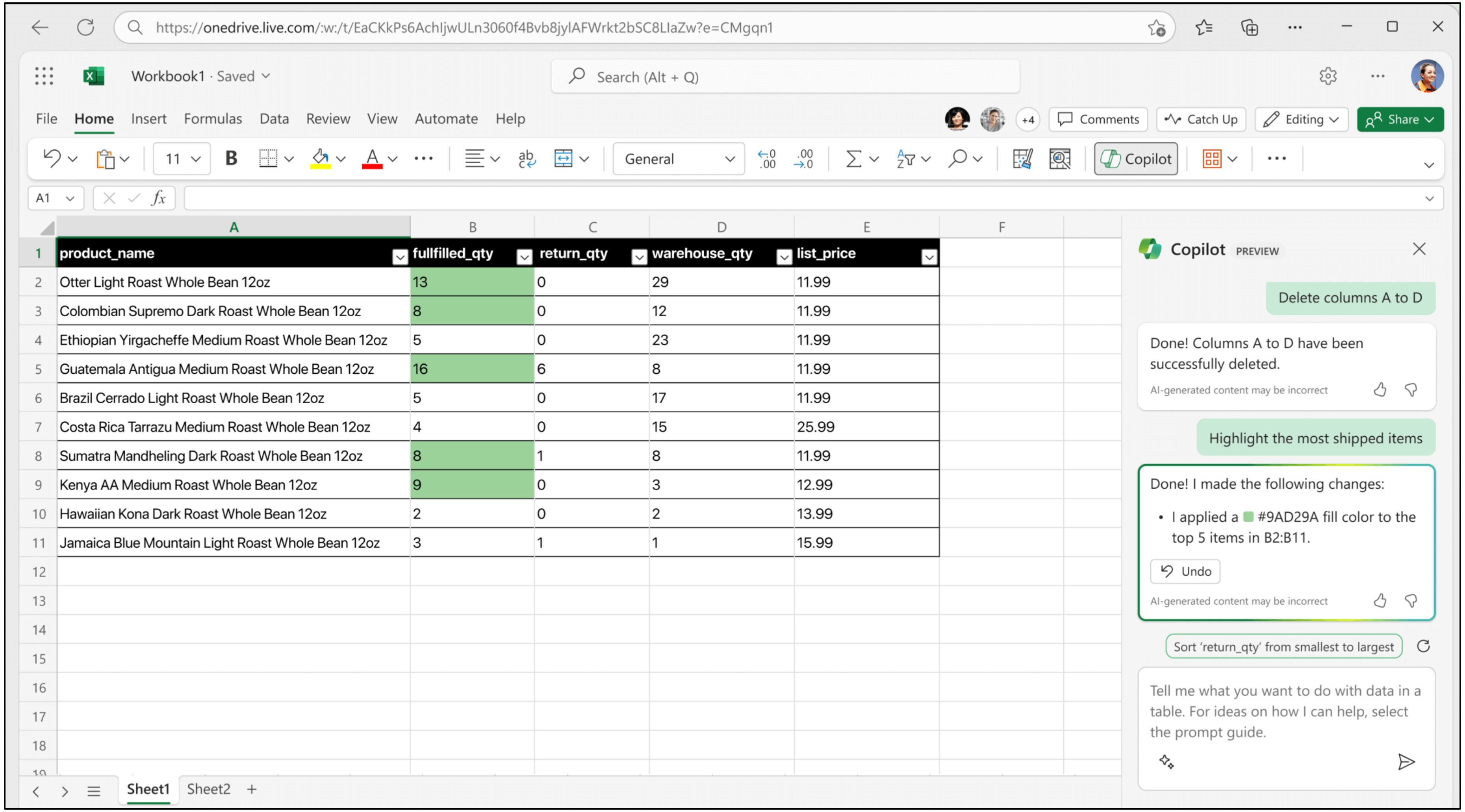
Alpine Ski House (sample)
someone8@example.com August 10, 2019 8:00 PM

A. Datum Corporation (sample)
someone9@example.com August 10, 2019 8:00 PM

Coho Winery (sample)
someone10@example.com August 10, 2019 8:00 PM

Fill
Background image None
Image position Fit
Loading spinner None

Ok, we shipped... how do we know if it is helpful?



product_name	fullfilled_qty	return_qty	warehouse_qty	list_price
Otter Light Roast Whole Bean 12oz	13	0	29	11.99
Colombian Supremo Dark Roast Whole Bean 12oz	8	0	12	11.99
Ethiopian Yirgacheffe Medium Roast Whole Bean 12oz	5	0	23	11.99
Guatemala Antigua Medium Roast Whole Bean 12oz	16	6	8	11.99
Brazil Cerrado Light Roast Whole Bean 12oz	5	0	17	11.99
Costa Rica Tarrazu Medium Roast Whole Bean 12oz	4	0	15	25.99
Sumatra Mandheling Dark Roast Whole Bean 12oz	8	1	8	11.99
Kenya AA Medium Roast Whole Bean 12oz	9	0	3	12.99
Hawaiian Kona Dark Roast Whole Bean 12oz	2	0	2	13.99
Jamaica Blue Mountain Light Roast Whole Bean 12oz	3	1	1	15.99

Copilot PREVIEW

Delete columns A to D

Done! Columns A to D have been successfully deleted.

AI-generated content may be incorrect

Highlight the most shipped items

Done! I made the following changes:

- I applied a 🟩 #9AD29A fill color to the top 5 items in B2:B11.

Undo

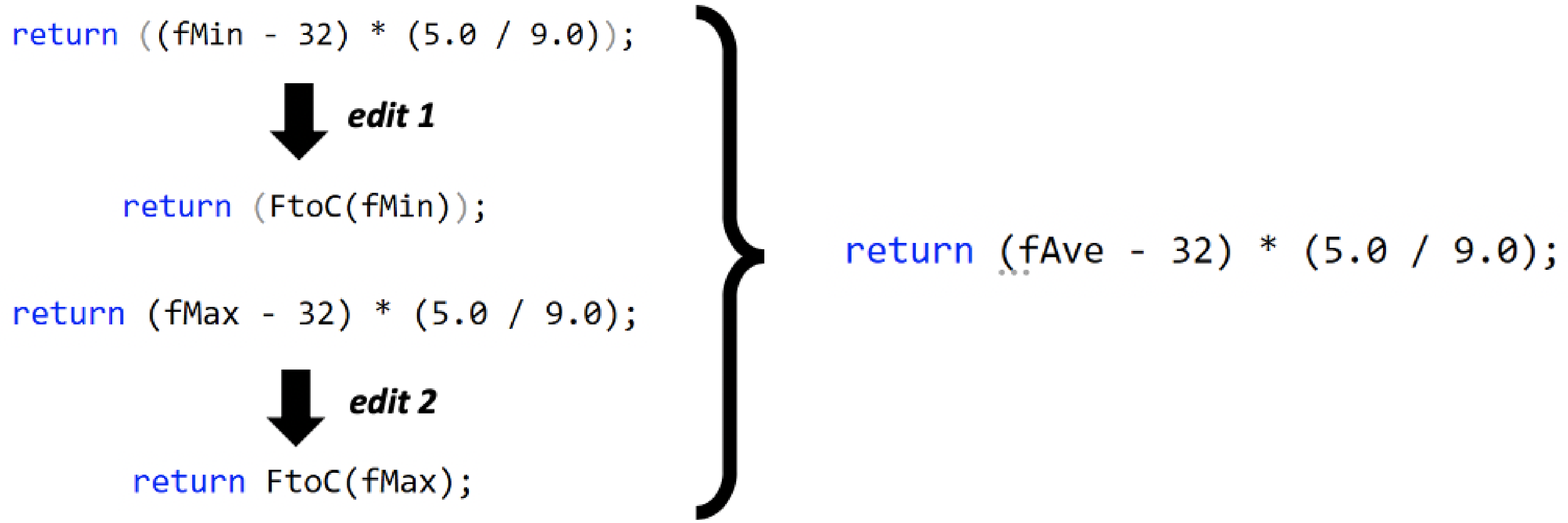
AI-generated content may be incorrect

Sort ‘return_qty’ from smallest to largest

Tell me what you want to do with data in a table. For ideas on how I can help, select the prompt guide.

How do we know when Excel Copilot is helping users? Not helping?

Let’s try 50 different UIs



```
return ((fMin - 32) * (5.0 / 9.0));
```

edit 1

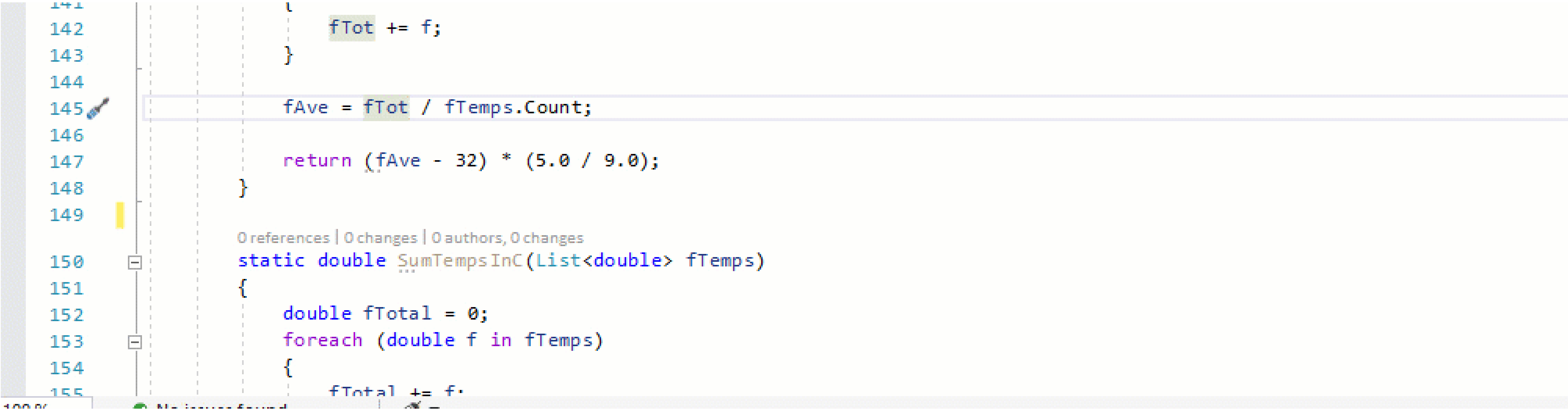
```
return (FtoC(fMin));
```

```
return (fMax - 32) * (5.0 / 9.0);
```

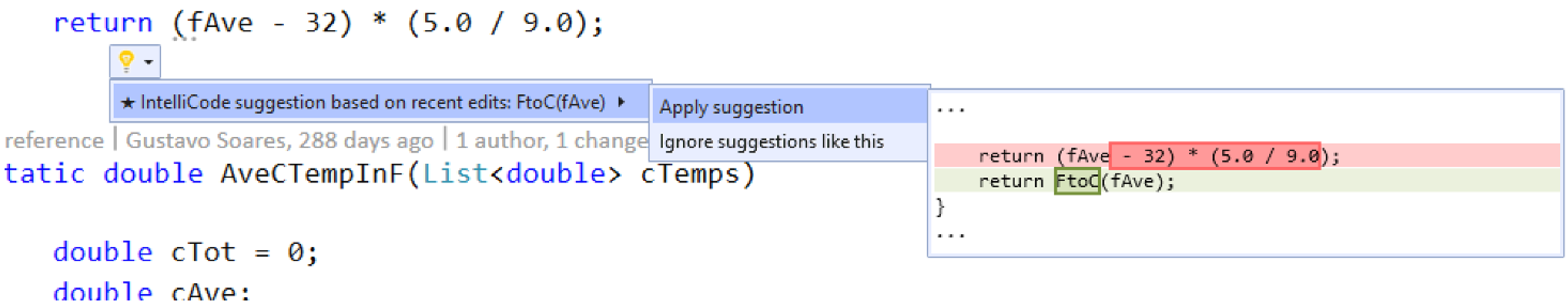
edit 2

```
return FtoC(fMax);
```

```
return (fAve - 32) * (5.0 / 9.0);
```



```
        {
            fTot += f;
        }

        fAve = fTot / fTemps.Count;

        return (fAve - 32) * (5.0 / 9.0);
    }

    0 references | 0 changes | 0 authors, 0 changes
    static double SumTempsInC(List<double> fTemps)
    {
        double fTotal = 0;
        foreach (double f in fTemps)
        {
            fTotal += f;
```

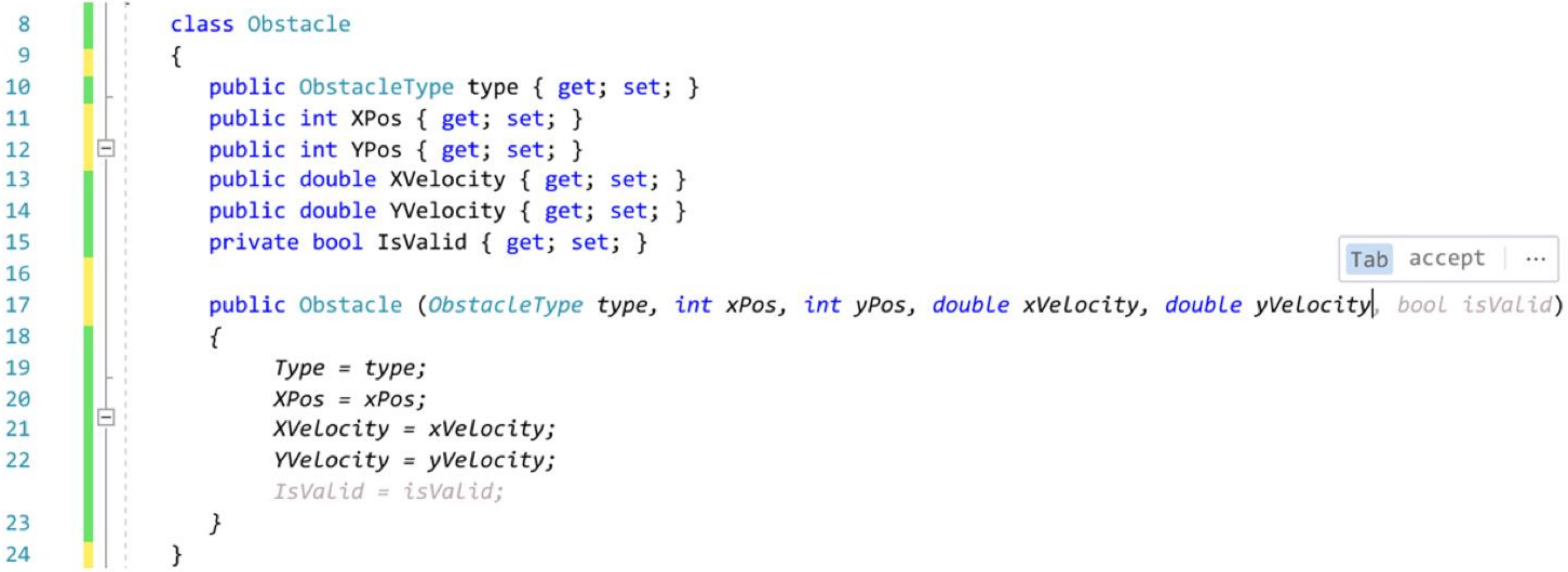


```
    return (fAve - 32) * (5.0 / 9.0);
```

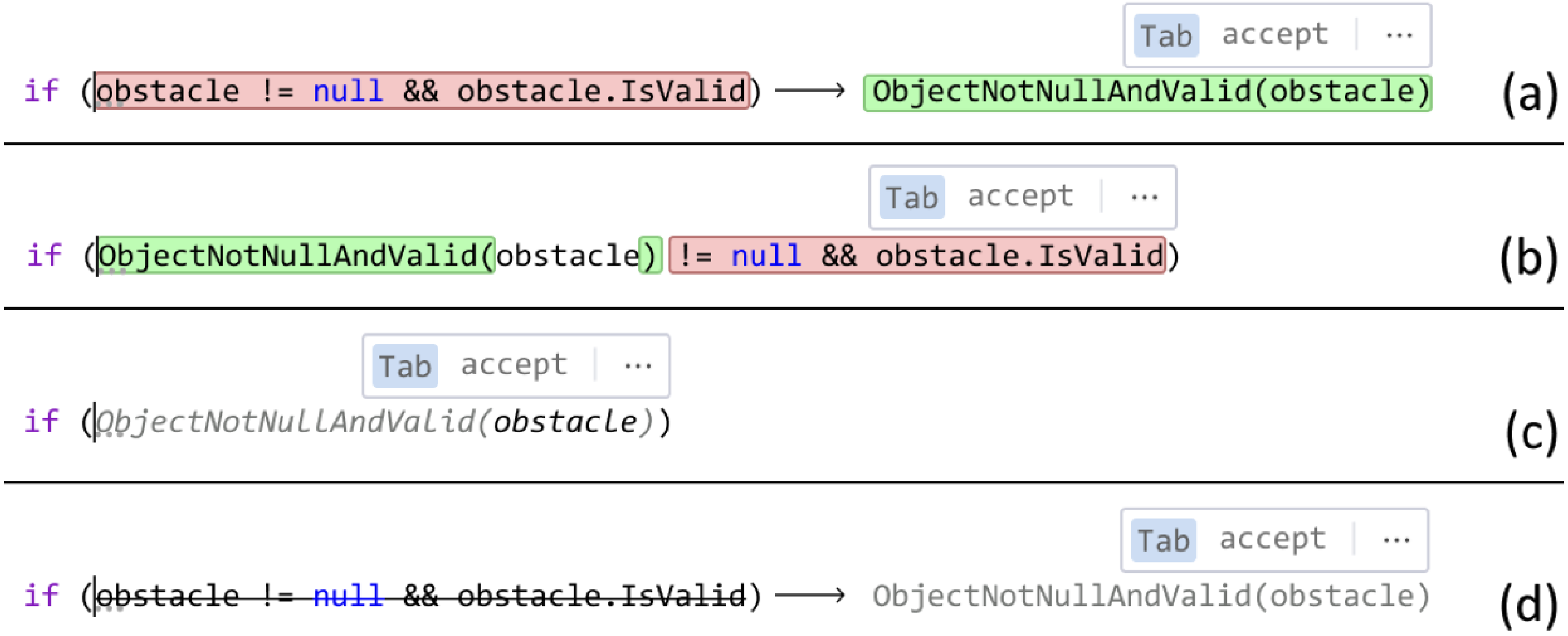
★ IntelliCode suggestion based on recent edits: FtoC(fAve)
Apply suggestion
Ignore suggestions like this
reference | Gustavo Soares, 288 days ago | 1 author, 1 change

```
tatic double AveCTempInF(List<double> cTemps)

    double cTot = 0;
    double cAve;
```

```
...

    return (fAve - 32) * (5.0 / 9.0);
    return FtoC(fAve);
}
...
```



```
class Obstacle
{
    public ObstacleType type { get; set; }
    public int XPos { get; set; }
    public int YPos { get; set; }
    public double XVelocity { get; set; }
    public double YVelocity { get; set; }
    private bool IsValid { get; set; }

    public Obstacle (ObstacleType type, int xPos, int yPos, double xVelocity, double yVelocity, bool isValid)
    {
        Type = type;
        XPos = xPos;
        XVelocity = xVelocity;
        YVelocity = yVelocity;
        IsValid = isValid;
    }
}
```

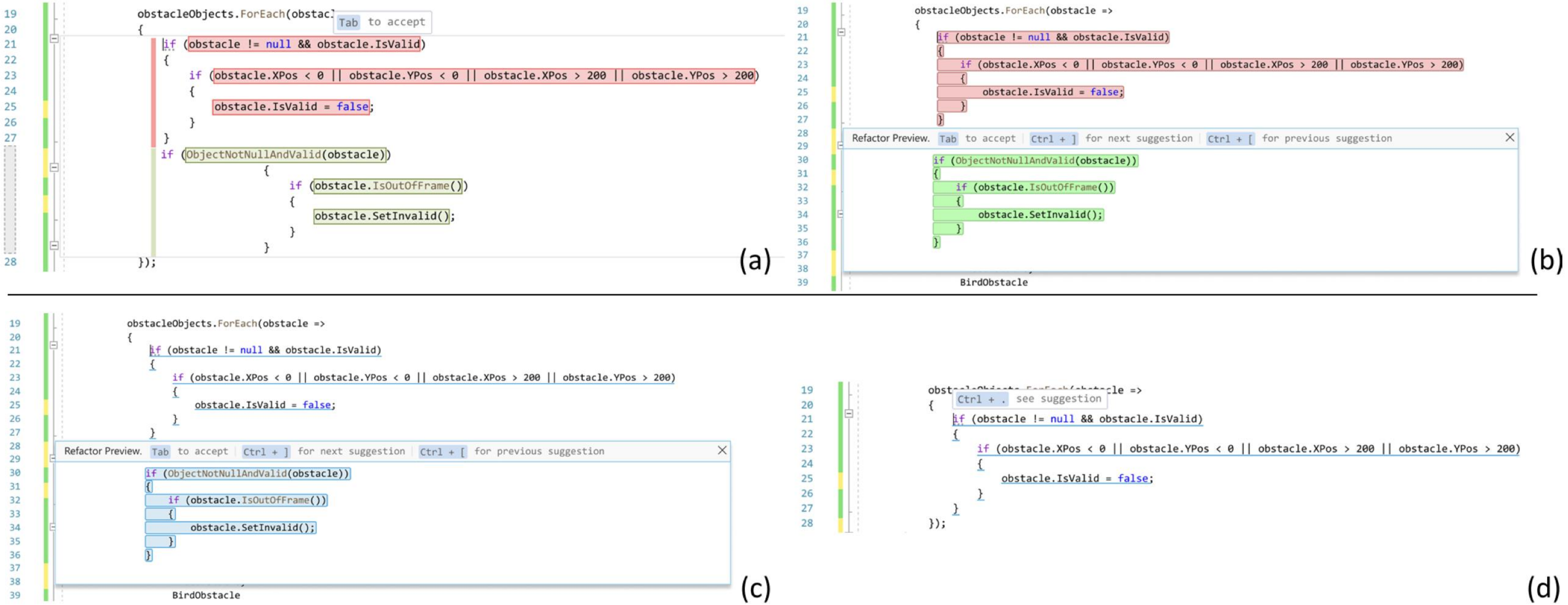


if (obstacle != null && obstacle.IsValid) ⟶ ObjectNotNullAndValid(obstacle)

if (ObjectNotNullAndValid(obstacle) != null && obstacle.IsValid)

if (ObjectNotNullAndValid(obstacle))

if (~~obstacle != null && obstacle.IsValid~~) ⟶ ObjectNotNullAndValid(obstacle)



(a)

```
obstacleObjects.ForEach(obstacl
{
    if (obstacle != null && obstacle.IsValid)
    {
        if (obstacle.XPos < 0 || obstacle.YPos < 0 || obstacle.XPos > 200 || obstacle.YPos > 200)
        {
            obstacle.IsValid = false;
        }
    }
    if (ObjectNotNullAndValid(obstacle))
            {
                if (obstacle.IsOutOfFrame())
                {
                    obstacle.SetInvalid();
                }
            }
});
```

Tab to accept

(b)

```
obstacleObjects.ForEach(obstacle =>
{
    if (obstacle != null && obstacle.IsValid)
    {
        if (obstacle.XPos < 0 || obstacle.YPos < 0 || obstacle.XPos > 200 || obstacle.YPos > 200)
        {
            obstacle.IsValid = false;
        }
    }
```

Refactor Preview. Tab to accept Ctrl +] for next suggestion Ctrl + [for previous suggestion

```
if (ObjectNotNullAndValid(obstacle))
{
    if (obstacle.IsOutOfFrame())
    {
        obstacle.SetInvalid();
    }
}
```

```
        BirdObstacle
```

(c)

```
obstacleObjects.ForEach(obstacle =>
{
    if (obstacle != null && obstacle.IsValid)
    {
        if (obstacle.XPos < 0 || obstacle.YPos < 0 || obstacle.XPos > 200 || obstacle.YPos > 200)
        {
            obstacle.IsValid = false;
        }
    }
```

Refactor Preview. Tab to accept Ctrl +] for next suggestion Ctrl + [for previous suggestion

```
if (ObjectNotNullAndValid(obstacle))
{
    if (obstacle.IsOutOfFrame())
    {
        obstacle.SetInvalid();
    }
}
```

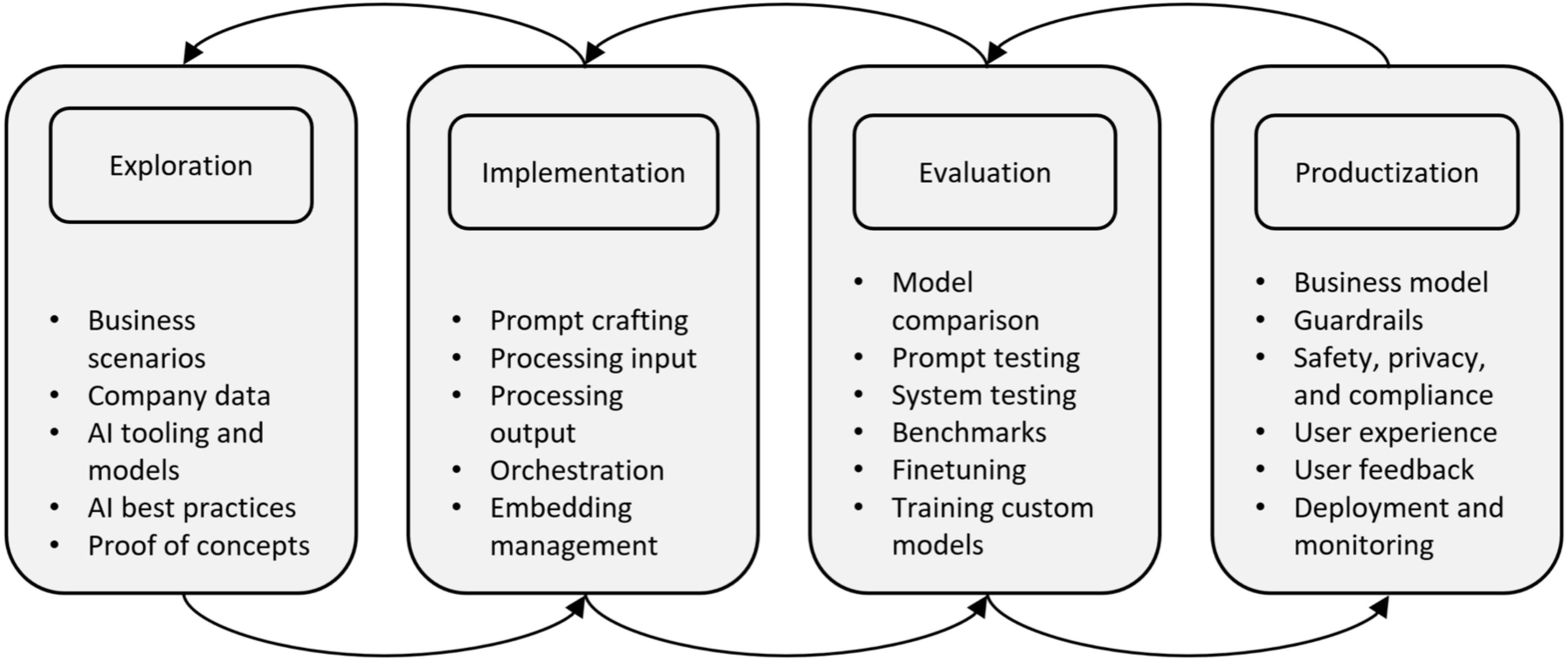
```
        BirdObstacle
```

(d)

Ctrl + . see suggestion

```
obst                    cle =>
{
    if (obstacle != null && obstacle.IsValid)
    {
        if (obstacle.XPos < 0 || obstacle.YPos < 0 || obstacle.XPos > 200 || obstacle.YPos > 200)
        {
            obstacle.IsValid = false;
        }
    }
});
```

How do we enable other developers to ship AI?



Exploration

- Business scenarios
- Company data
- AI tooling and models
- AI best practices
- Proof of concepts

Implementation

- Prompt crafting
- Processing input
- Processing output
- Orchestration
- Embedding management

Evaluation

- Model comparison
- Prompt testing
- System testing
- Benchmarks
- Finetuning
- Training custom models

Productization

- Business model
- Guardrails
- Safety, privacy, and compliance
- User experience
- User feedback
- Deployment and monitoring

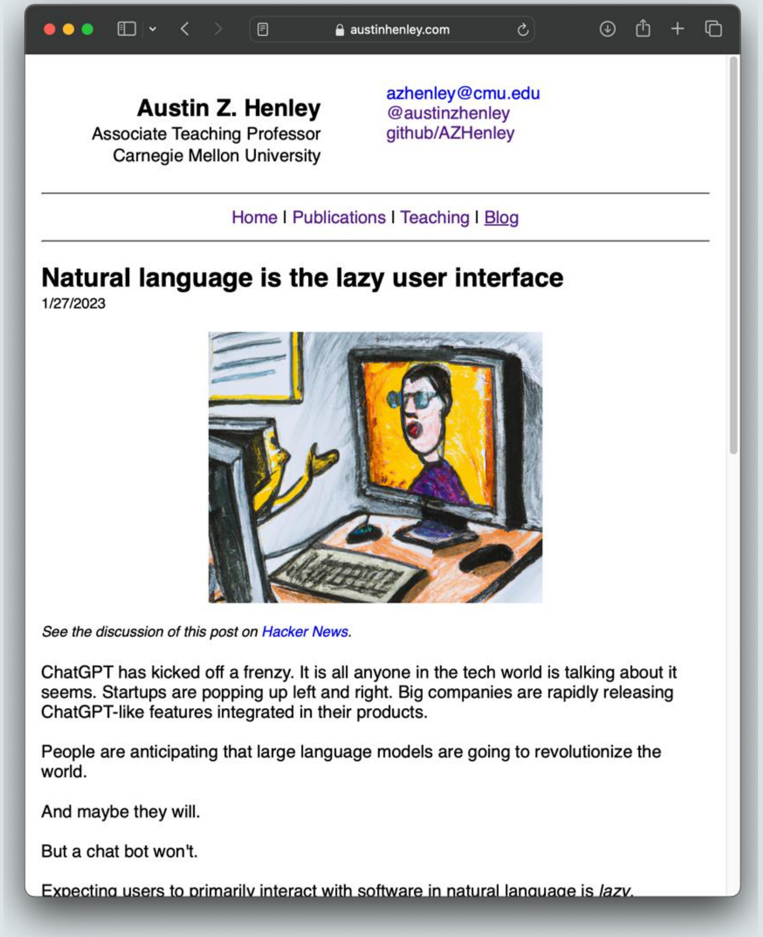
People were using unit tests to...

Lessons, maybe

Are you a solution looking for a problem?

Big companies have many constraints—many have nothing to do with how good your tech is

Innovator’s dilemma?



Austin Z. Henley
Associate Teaching Professor
Carnegie Mellon University

azhenley@cmu.edu
@austinzhenley
github/AZHenley

Home | Publications | Teaching | Blog

Natural language is the lazy user interface

1/27/2023

See the discussion of this post on Hacker News.

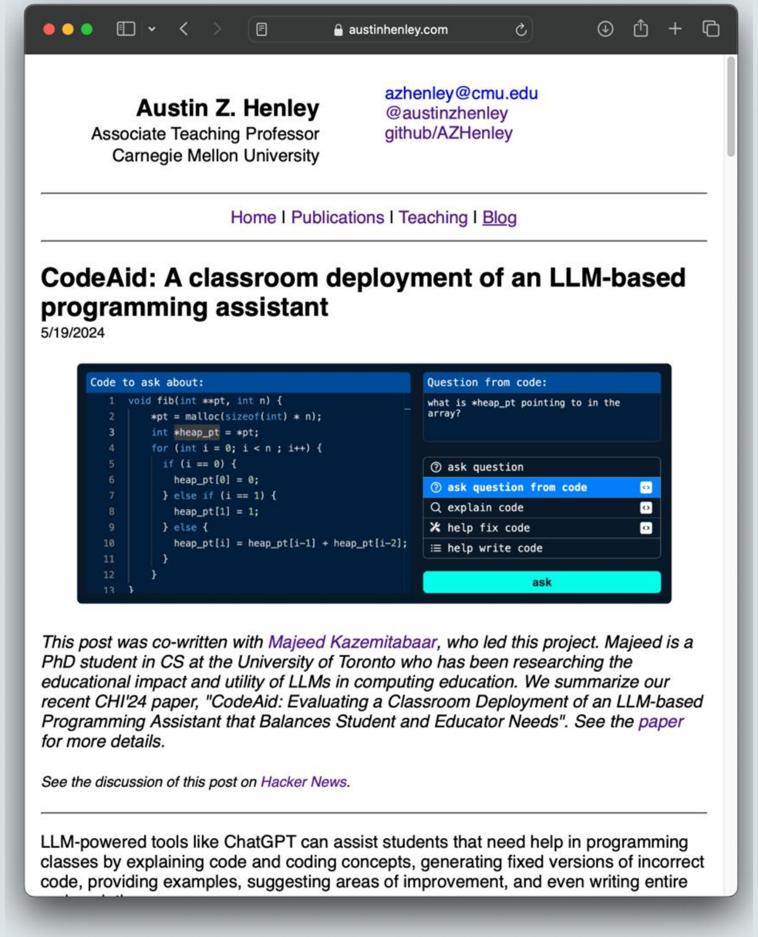
ChatGPT has kicked off a frenzy. It is all anyone in the tech world is talking about it seems. Startups are popping up left and right. Big companies are rapidly releasing ChatGPT-like features integrated in their products.

People are anticipating that large language models are going to revolutionize the world.

And maybe they will.

But a chat bot won't.

Expecting users to primarily interact with software in natural language is *lazy*.



Austin Z. Henley
Associate Teaching Professor
Carnegie Mellon University

azhenley@cmu.edu
@austinzhenley
github/AZHenley

Home | Publications | Teaching | Blog

CodeAid: A classroom deployment of an LLM-based programming assistant

5/19/2024

Code to ask about:

```
void fib(int **pt, int n) {
    *pt = malloc(sizeof(int) * n);
    int *heap_pt = *pt;
    for (int i = 0; i < n ; i++) {
      if (i == 0) {
        heap_pt[0] = 0;
      } else if (i == 1) {
        heap_pt[1] = 1;
      } else {
        heap_pt[i] = heap_pt[i-1] + heap_pt[i-2];
      }
    }
}
```

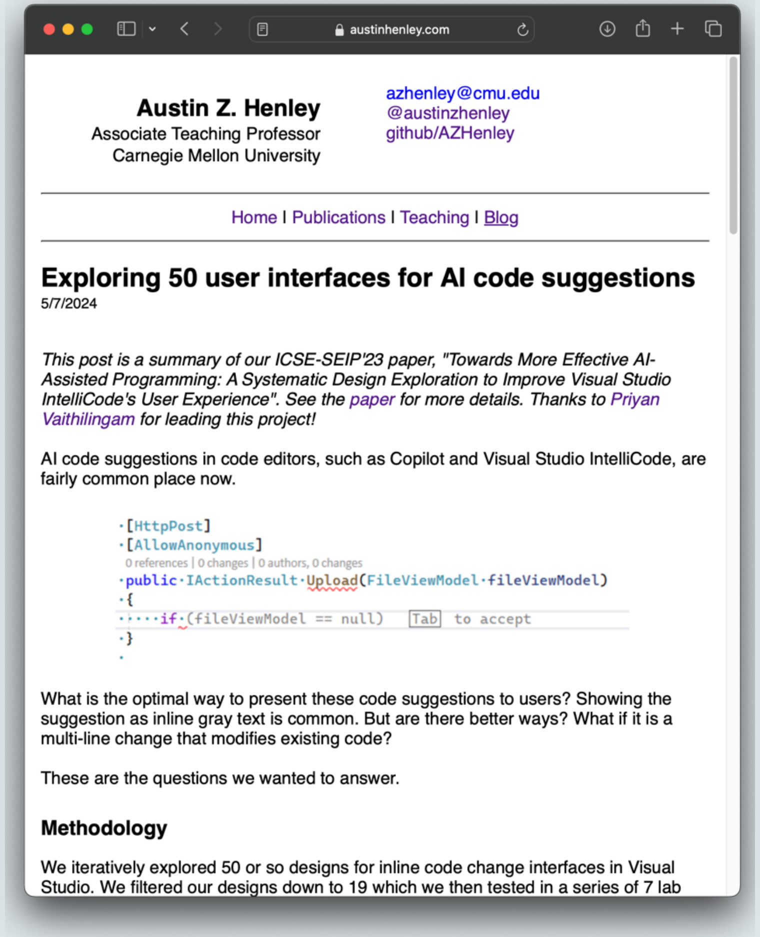
Question from code:

what is *heap_pt pointing to in the array?

This post was co-written with Majeed Kazemitabaar, who led this project. Majeed is a PhD student in CS at the University of Toronto who has been researching the educational impact and utility of LLMs in computing education. We summarize our recent CHI'24 paper, "CodeAid: Evaluating a Classroom Deployment of an LLM-based Programming Assistant that Balances Student and Educator Needs". See the paper for more details.

See the discussion of this post on Hacker News.

LLM-powered tools like ChatGPT can assist students that need help in programming classes by explaining code and coding concepts, generating fixed versions of incorrect code, providing examples, suggesting areas of improvement, and even writing entire



Austin Z. Henley
Associate Teaching Professor
Carnegie Mellon University

azhenley@cmu.edu
@austinzhenley
github/AZHenley

Home | Publications | Teaching | Blog

Exploring 50 user interfaces for AI code suggestions

5/7/2024

This post is a summary of our ICSE-SEIP'23 paper, "Towards More Effective AI-Assisted Programming: A Systematic Design Exploration to Improve Visual Studio IntelliCode's User Experience". See the paper for more details. Thanks to Priyan Vaithilingam for leading this project!

AI code suggestions in code editors, such as Copilot and Visual Studio IntelliCode, are fairly common place now.

```
·[HttpPost]
·[AllowAnonymous]
 0 references | 0 changes | 0 authors, 0 changes
·public·IActionResult·Upload(FileViewModel·fileViewModel)
·{
·····if·(fileViewModel == null)   Tab  to accept
·}
·
```

What is the optimal way to present these code suggestions to users? Showing the suggestion as inline gray text is common. But are there better ways? What if it is a multi-line change that modifies existing code?

These are the questions we wanted to answer.

Methodology

We iteratively explored 50 or so designs for inline code change interfaces in Visual Studio. We filtered our designs down to 19 which we then tested in a series of 7 lab