NCAIR 2024 Annual Conference

Institutional Innovation

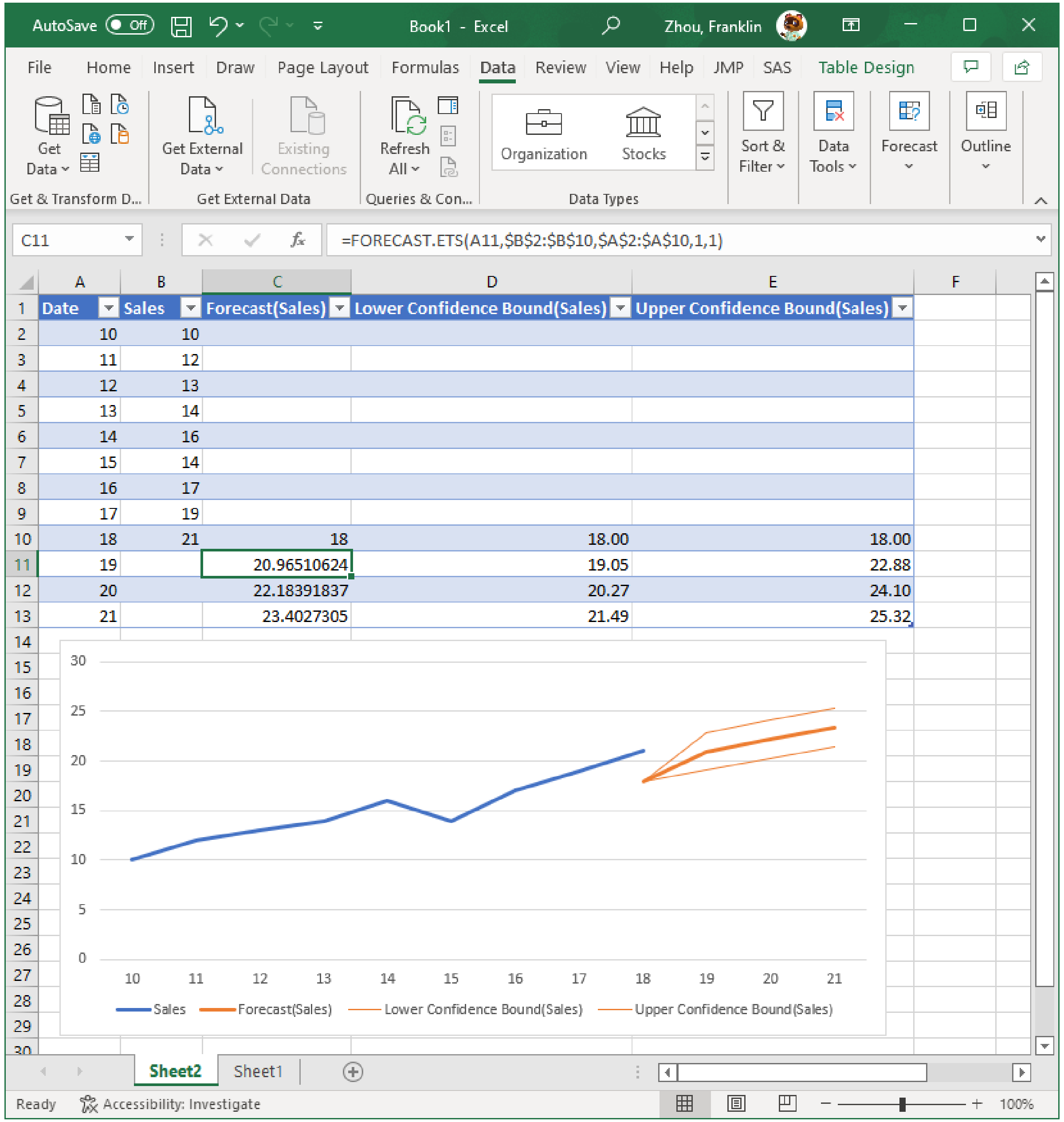
Statistic Tools Introduction

Franklin Zhou
IPAR, East Carolina University

- Excel
- R
- SAS
- Python
- JMP
- SPSS
- Minitab
- …

Excel

- One of the most widely used software
- Up to 1 million rows and 16,000 columns per worksheet.
- Can create and customize charts, graphs, tables, and pivot tables to visualize and summarize data
- Includes many useful built-in functions.
- Limited in functionality and flexibility



=FORECAST.ETS(A11,$B$2:$B$10,$A$2:$A$10,1,1)

Date	Sales	Forecast(Sales)	Lower Confidence Bound(Sales)	Upper Confidence Bound(Sales)
10	10			
11	12			
12	13			
13	14			
14	16			
15	14			
16	17			
17	19			
18	21	18	18.00	18.00
19		20.96510624	19.05	22.88
20		22.18391837	20.27	24.10
21		23.4027305	21.49	25.32

What appeals to me

- Easy to create and modify data
- Dynamic updates
- Font and Style
- Directly working on the data makes “WYSIWYG”
- Quick pivot table
- Copy & paste works well with other MS Office apps
- Collaborative work
- Be able to write and draw notes

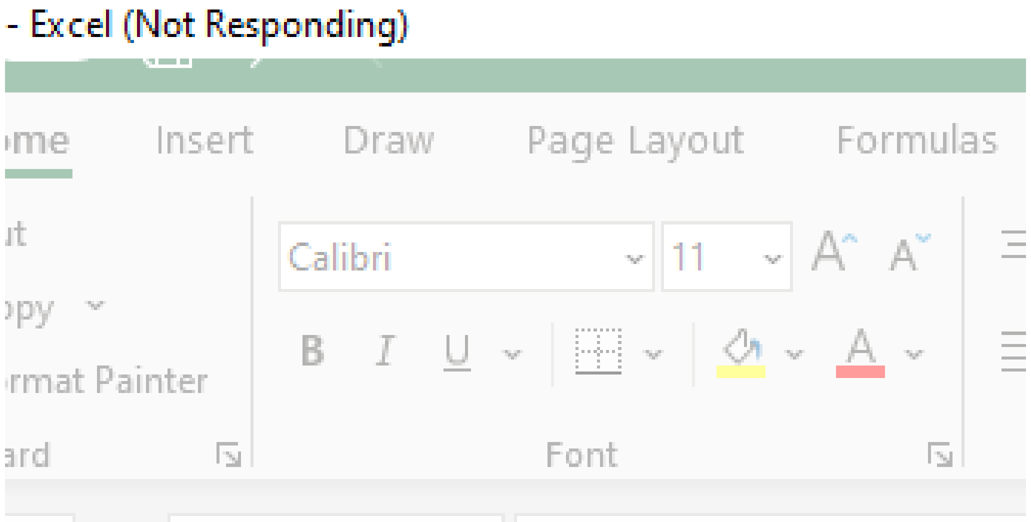
What troubles me

- Formats
- Customize charts
- Build models
- Not friendly with large dataset
- Undo

What appeals to me

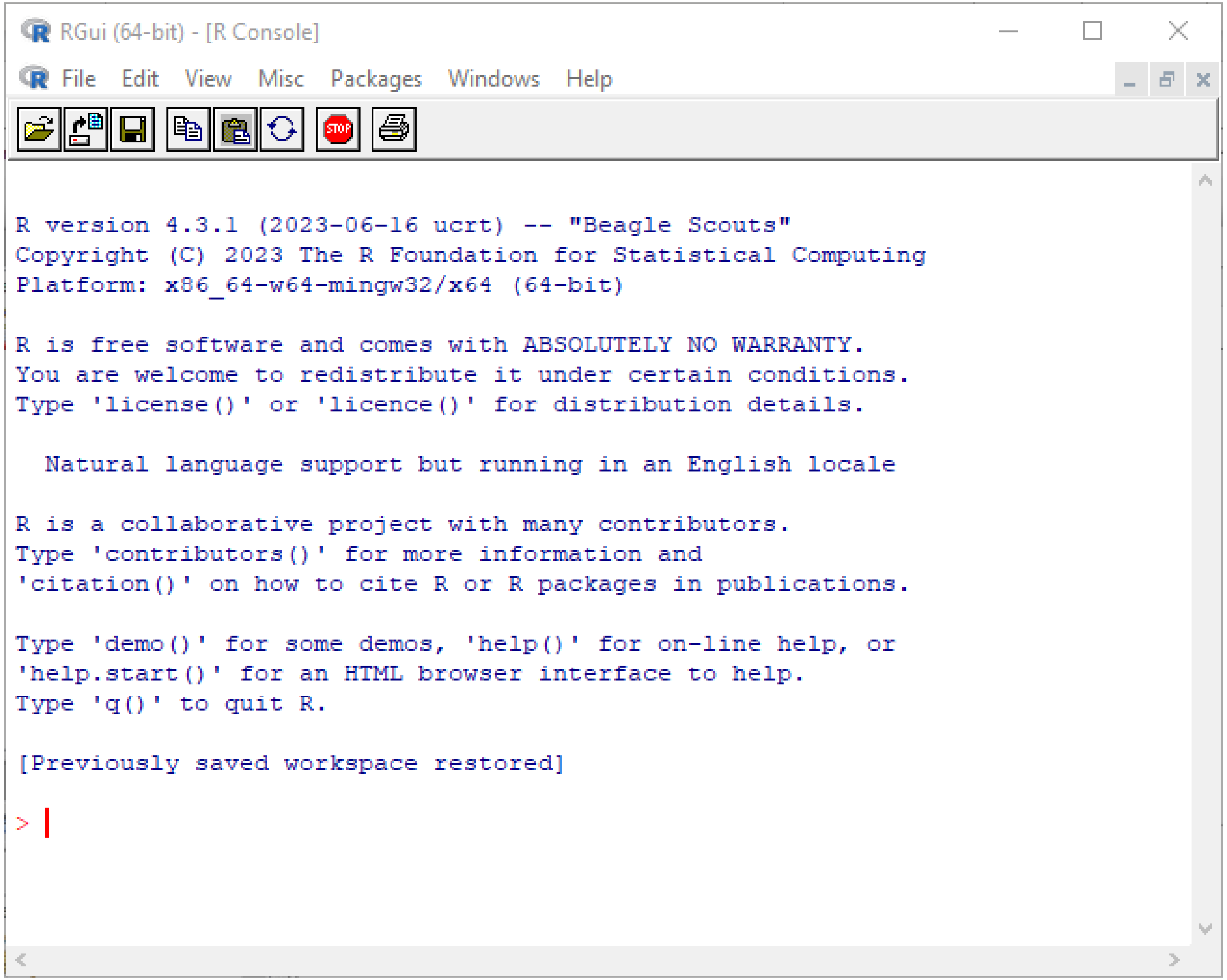
- Easy to create and modify data
- Dynamic updates
- Font and Style
- Directly working on the data makes “WYSIWYG”
- Quick pivot table
- Copy & paste works well with other MS Office apps
- Collaborative work
- Be able to write and draw notes

What troubles me



- Excel (Not Responding)

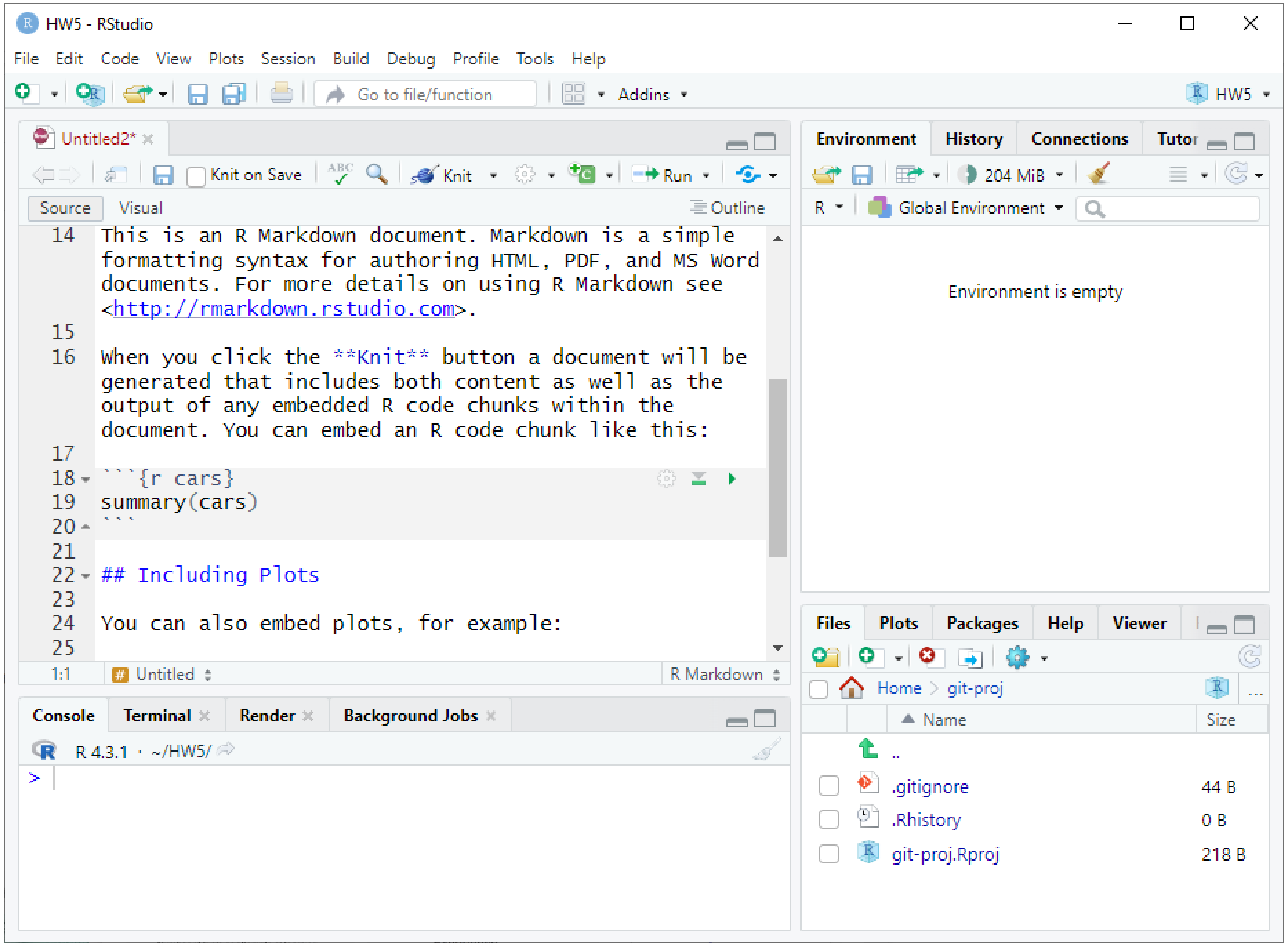
- One of the most popular **free** statistics software
- Programming-based
- Rely on third-party contributions
- Lots of online resources


RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

```
R version 4.3.1 (2023-06-16 ucrt) -- "Beagle Scouts"
Copyright (C) 2023 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.

[Previously saved workspace restored]

> 
```

RStudio®

- **Free** tool that provides integrated development environment(IDE) to use R language
- Resource management
- Programming friendly
- More than R script:
    - R Quarto/Markdown notebook and presentations
    - R Shiny Web App



~~~
This is an R Markdown document. Markdown is a simple
formatting syntax for authoring HTML, PDF, and MS Word
documents. For more details on using R Markdown see
<http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be
generated that includes both content as well as the
output of any embedded R code chunks within the
document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```

## Including Plots

You can also embed plots, for example:
~~~

Environment is empty

Name	Size
..	
.gitignore	44 B
.Rhistory	0 B
git-proj.Rproj	218 B

```
R 4.3.1 · ~/HW5/
>
```

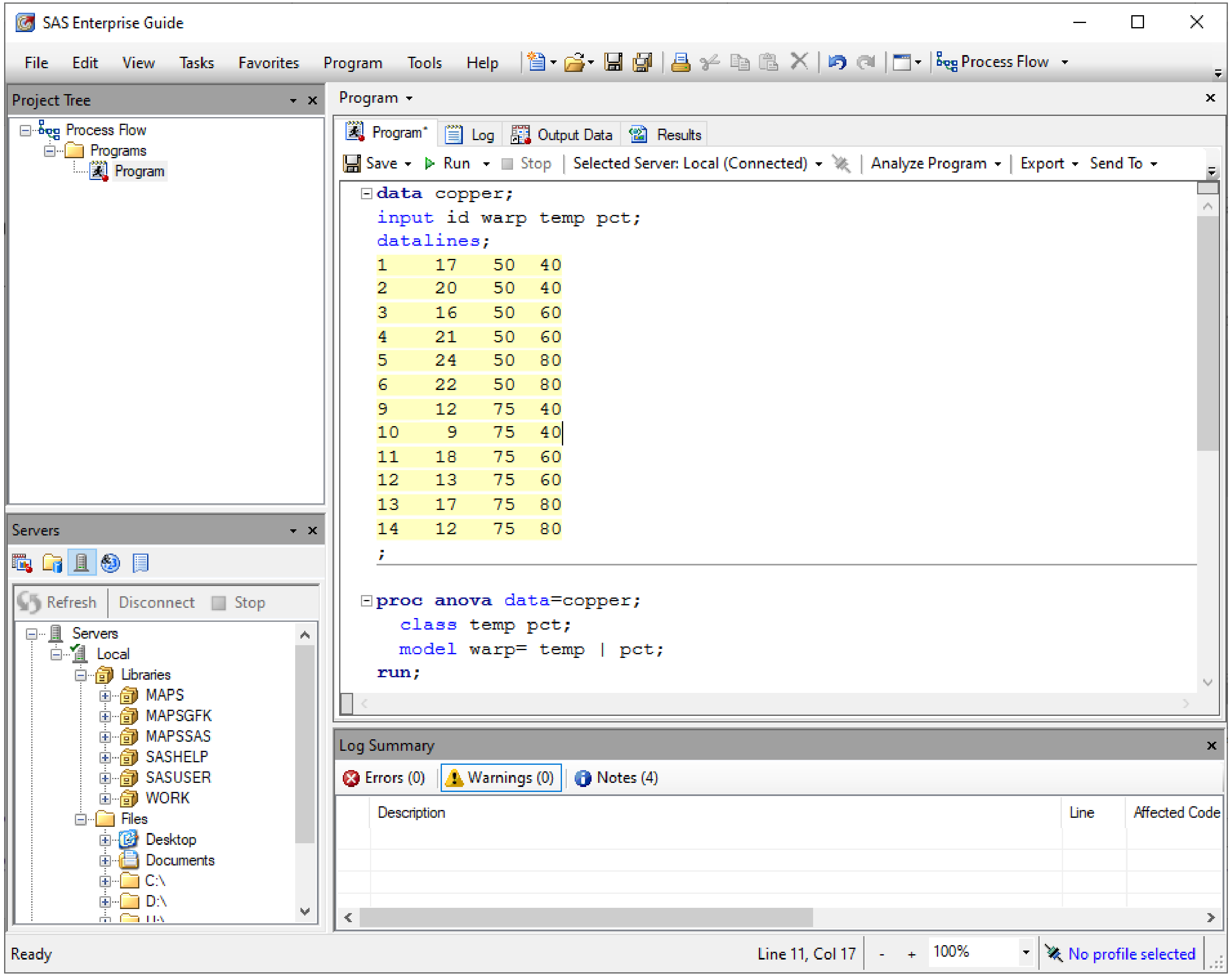
What appeals to me

- Codes can be edited and reused
- More functions than Excel
- Customize plots
- Project management
- Be able to link with GitHub*

What troubles me

- Not as easy as Excel to view and edit data
- Too many ways to solve a problem
- Packages update issue
- Lack of “WYSIWYG”

- One of the most popular commercial data analytics software
- Programming-based
- All functions are built-in
- Lots of online resources



```
data copper;
input id warp temp pct;
datalines;
1    17   50  40
2    20   50  40
3    16   50  60
4    21   50  60
5    24   50  80
6    22   50  80
9    12   75  40
10    9   75  40
11   18   75  60
12   13   75  60
13   17   75  80
14   12   75  80
;

proc anova data=copper;
  class temp pct;
  model warp= temp | pct;
run;
```

What appeals to me

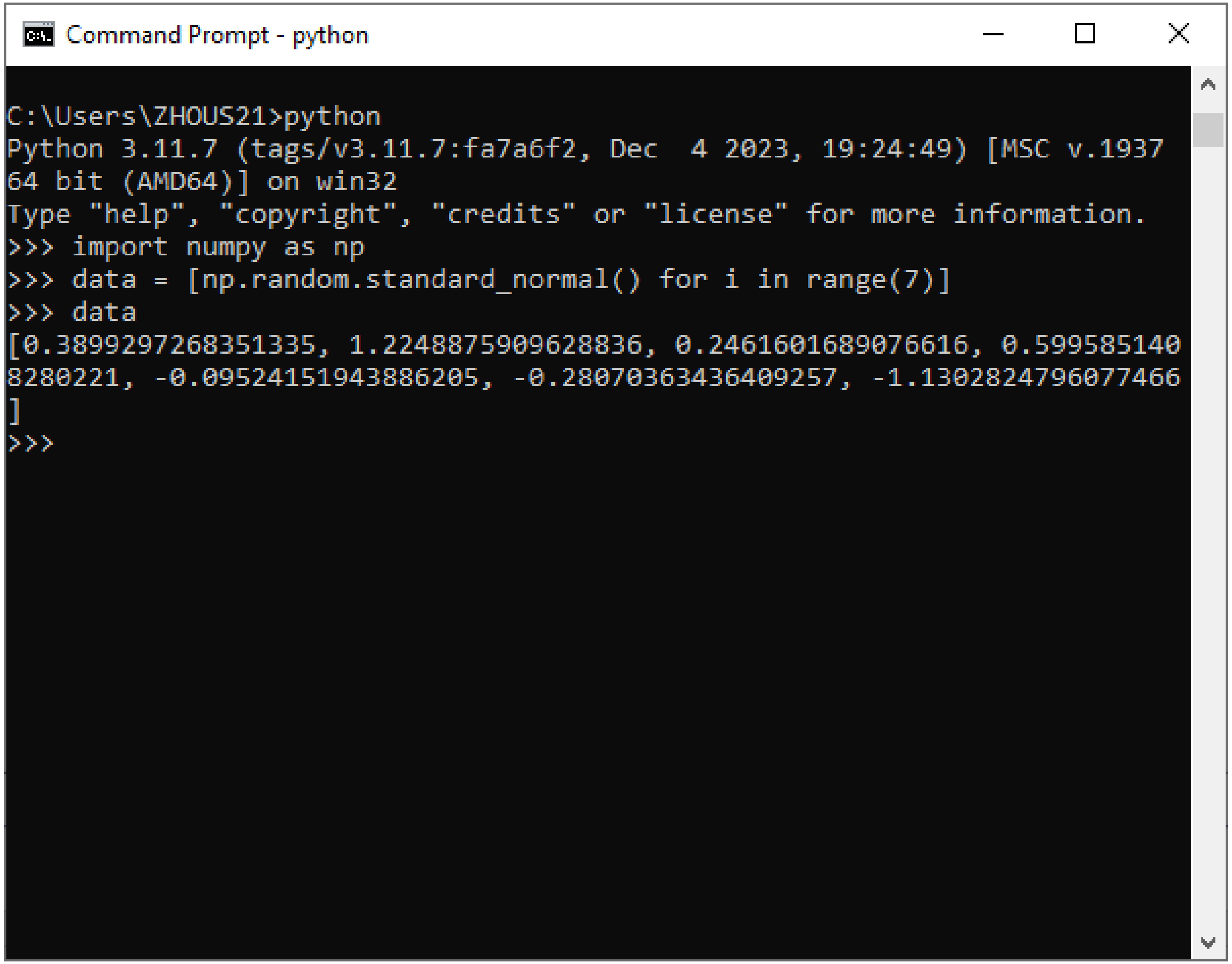
- Codes can be edited and reused
- Less flexible coding
- More functions than Excel
- Customize plots
- Project management

What troubles me

- Not free to use
- Not as easy as Excel to view and edit data
- Lack of “WYSIWYG”

python

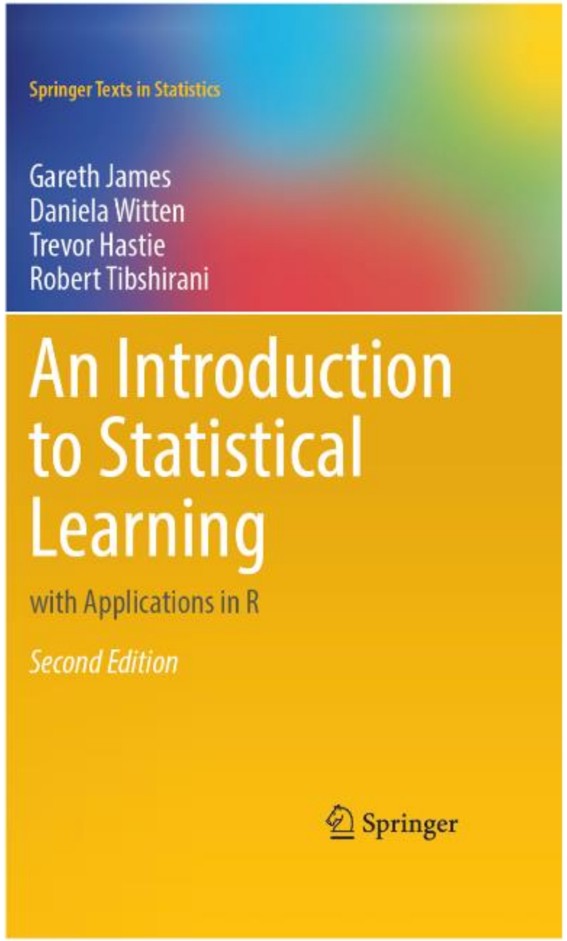
- Emerging competitor of R in data analysis field
- Has many same features as R:
  - Programming-based
  - Rely on third-party contributions
  - Lots of online resources



```
C:\Users\ZHOUS21>python
Python 3.11.7 (tags/v3.11.7:fa7a6f2, Dec  4 2023, 19:24:49) [MSC v.1937
64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import numpy as np
>>> data = [np.random.standard_normal() for i in range(7)]
>>> data
[0.3899297268351335, 1.2248875909628836, 0.2461601689076616, 0.599585140
8280221, -0.09524151943886205, -0.28070363436409257, -1.1302824796077466
]
>>>
```

python

- Emerging competitor of R in data analysis field
- Has many same features as R:
  - Programming-based
  - Rely on third-party contributions
  - Lots of online resources



Springer Texts in Statistics

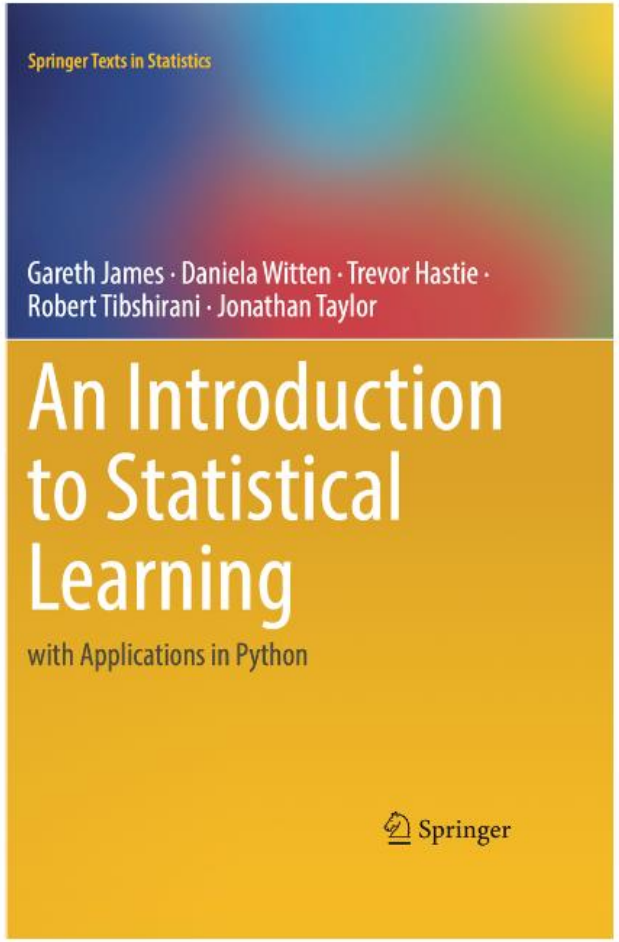
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Second Edition

Springer



Springer Texts in Statistics

Gareth James · Daniela Witten · Trevor Hastie ·
Robert Tibshirani · Jonathan Taylor

An Introduction to Statistical Learning

with Applications in Python

Springer

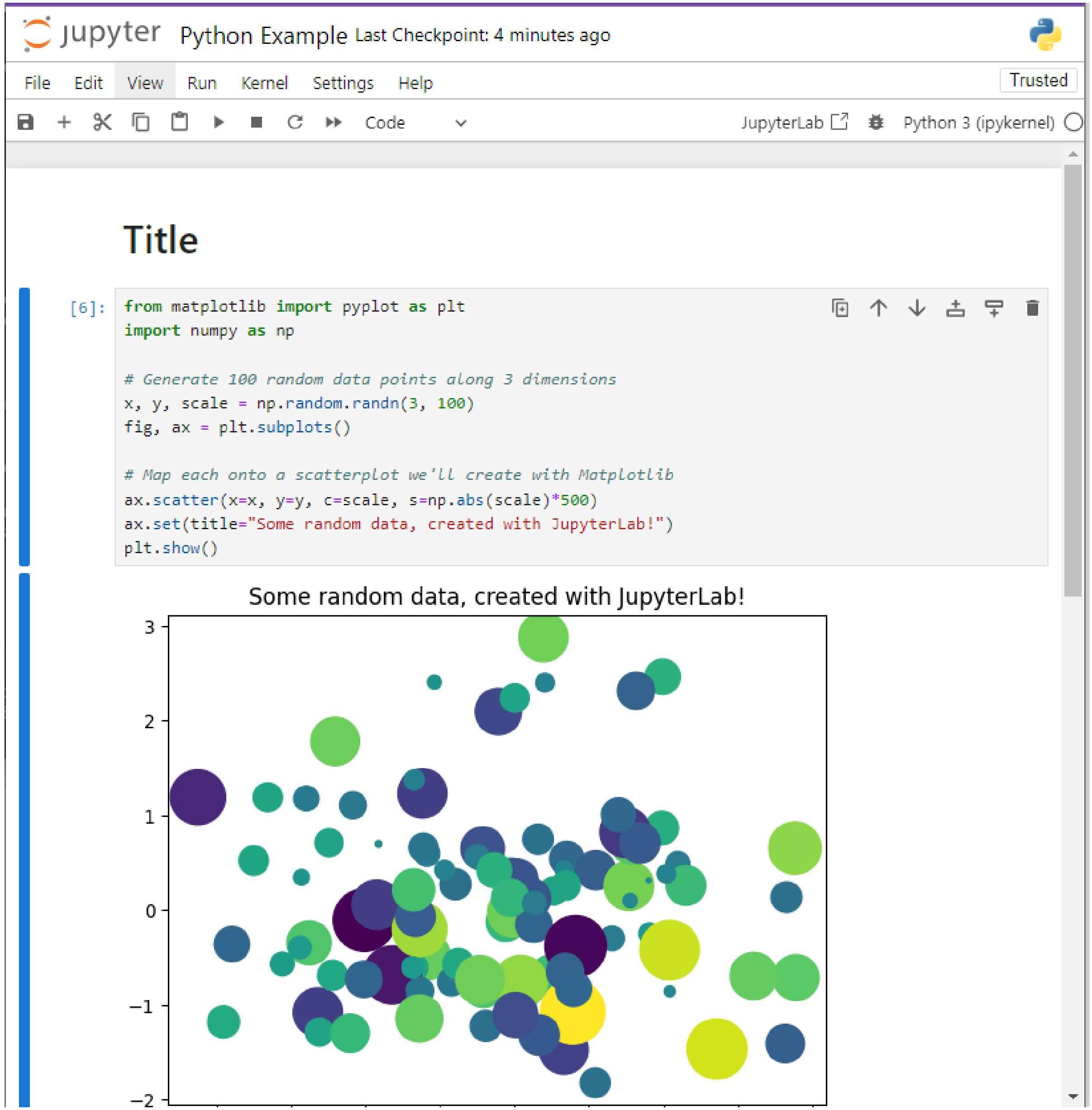
jupyter

- Markdown for Python
- Free

JupyterLab: A Next-Generation Notebook Interface

Jupyter Notebook: The Classic Notebook Interface

Voilà: Share your results (web applications)



```
from matplotlib import pyplot as plt
import numpy as np

# Generate 100 random data points along 3 dimensions
x, y, scale = np.random.randn(3, 100)
fig, ax = plt.subplots()

# Map each onto a scatterplot we'll create with Matplotlib
ax.scatter(x=x, y=y, c=scale, s=np.abs(scale)*500)
ax.set(title="Some random data, created with JupyterLab!")
plt.show()
```

A snapshot of Jupyter Notebook

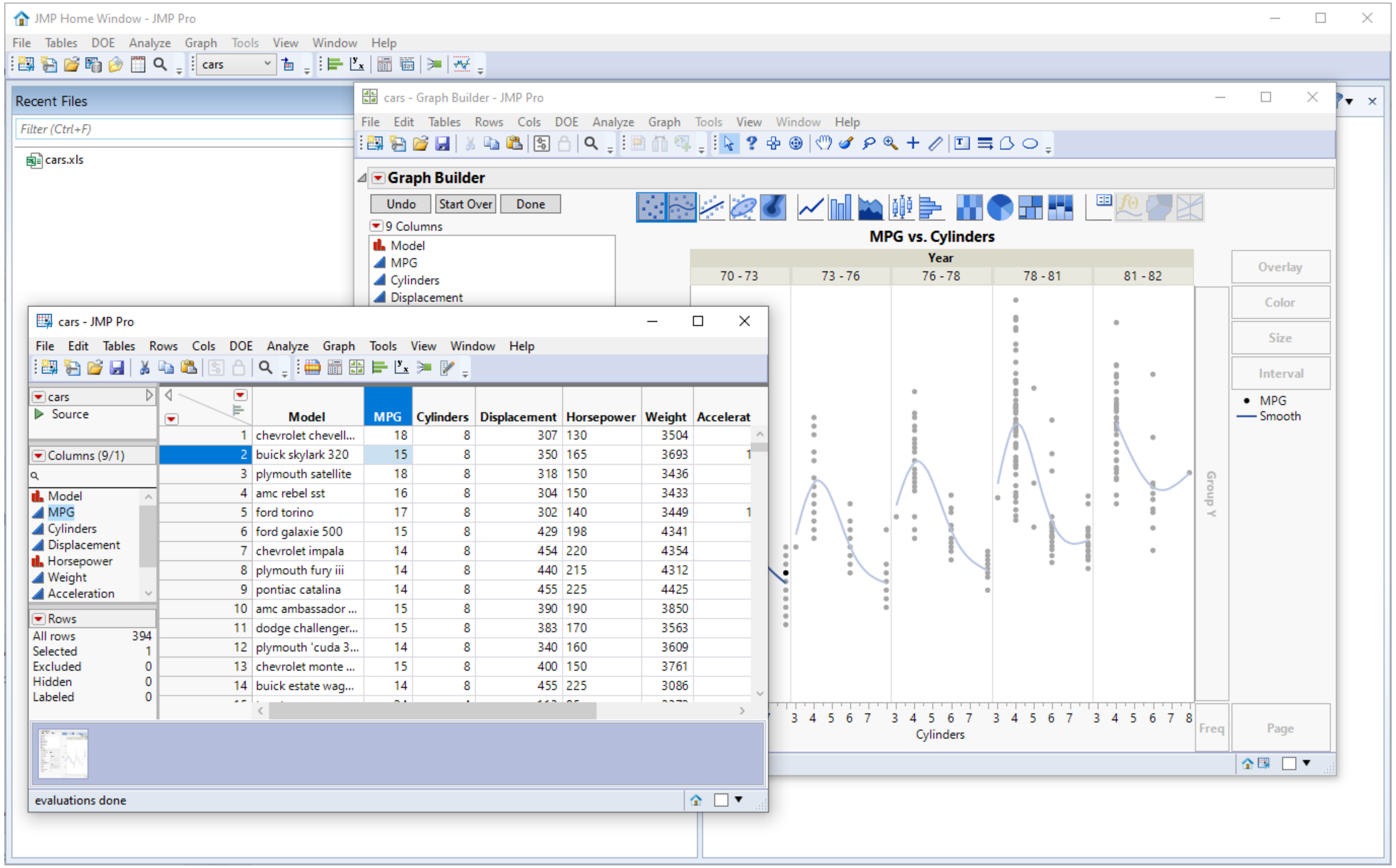
What appeals to me

- Similar as R and R Studio
- Has advantages in Deep Learning, Machine Learning, Data Mining, and AI integration.
- Can do more besides statistical analysis

What troubles me

- Not as easy as Excel to view and edit data
- Too many ways to solve a problem
- Version update issue
- Lack of “WYSIWYG”

- A product of SAS company
- Easy to use with GUI
- Good for plotting and building some advanced models



	Model	MPG	Cylinders	Displacement	Horsepower	Weight	Accelerat
1	chevrolet chevell...	18	8	307	130	3504	
2	buick skylark 320	15	8	350	165	3693	1
3	plymouth satellite	18	8	318	150	3436	
4	amc rebel sst	16	8	304	150	3433	
5	ford torino	17	8	302	140	3449	1
6	ford galaxie 500	15	8	429	198	4341	
7	chevrolet impala	14	8	454	220	4354	
8	plymouth fury iii	14	8	440	215	4312	
9	pontiac catalina	14	8	455	225	4425	
10	amc ambassador ...	15	8	390	190	3850	
11	dodge challenger...	15	8	383	170	3563	
12	plymouth 'cuda 3...	14	8	340	160	3609	
13	chevrolet monte ...	15	8	400	150	3761	
14	buick estate wag...	14	8	455	225	3086	

Rows	
All rows	394
Selected	1
Excluded	0
Hidden	0
Labeled	0

evaluations done

What appeals to me

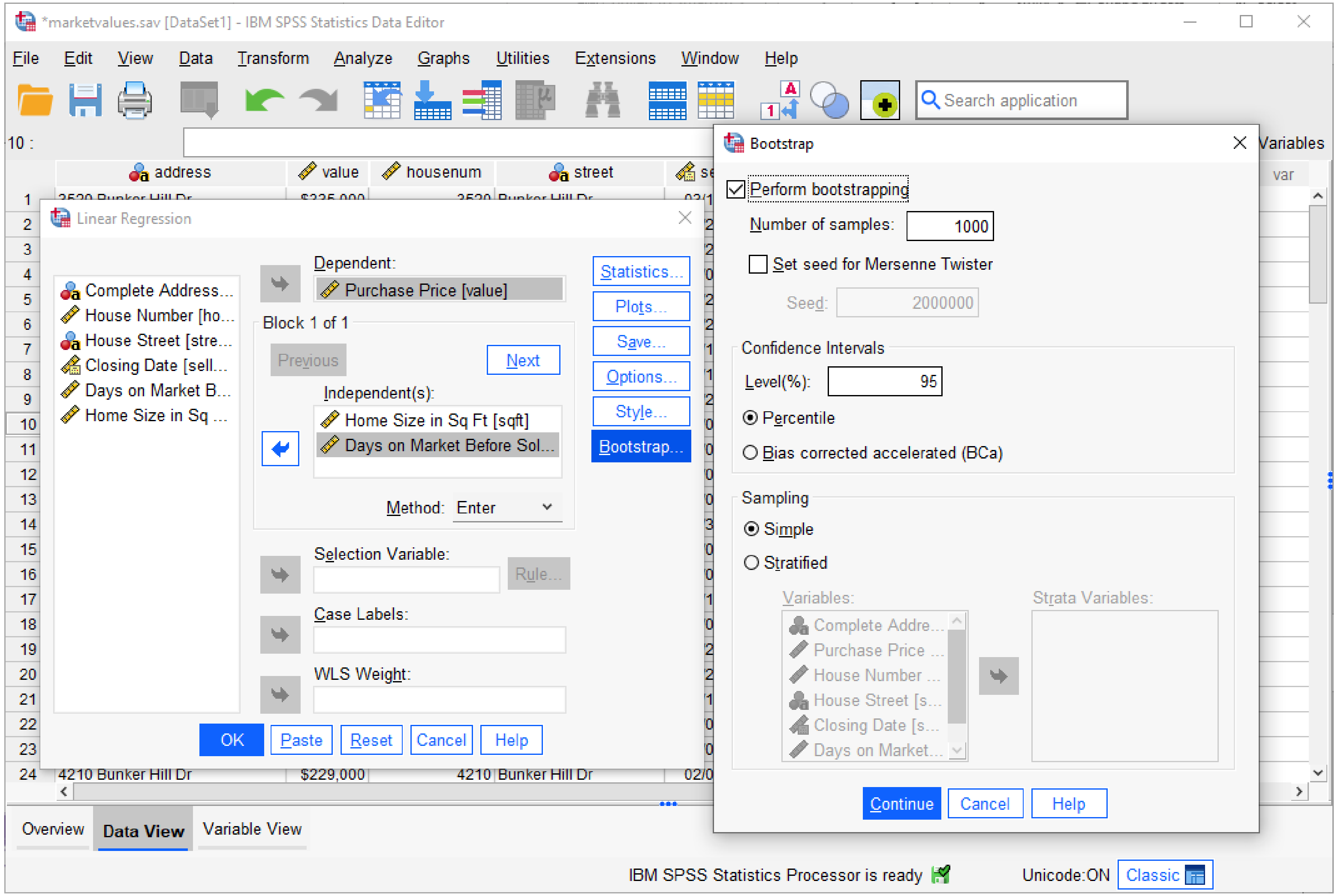
- Simple and quick of making figures/plots
- Simple and quick of building certain models:
  - PCA, SEM, Factor Analysis
  - Neural, KNN, SVM
  - Hierarchical Cluster, K Means
  - Nonlinear, Time Series
  - Survival Analysis
  - Text Mining
- JMP offers free webinar every week.

What troubles me

- Not be able to replicate all steps with one button.
- Not be able to cover all statistical methods.
- Not flexible as R / Python

SPSS

- A commercial product from IBM company
- Easy to use with GUI
- Can cover most statistical analysis works
- Supports syntax command



*marketvalues.sav [DataSet1] - IBM SPSS Statistics Data Editor

Linear Regression

Dependent: Purchase Price [value]

Block 1 of 1

Independent(s):
Home Size in Sq Ft [sqft]
Days on Market Before Sol...

Method: Enter

Bootstrap

☑ Perform bootstrapping

Number of samples: 1000

□ Set seed for Mersenne Twister

Seed: 2000000

Confidence Intervals

Level(%): 95

● Percentile

○ Bias corrected accelerated (BCa)

Sampling

● Simple

○ Stratified

IBM SPSS Statistics Processor is ready

Unicode:ON

What appeals to me

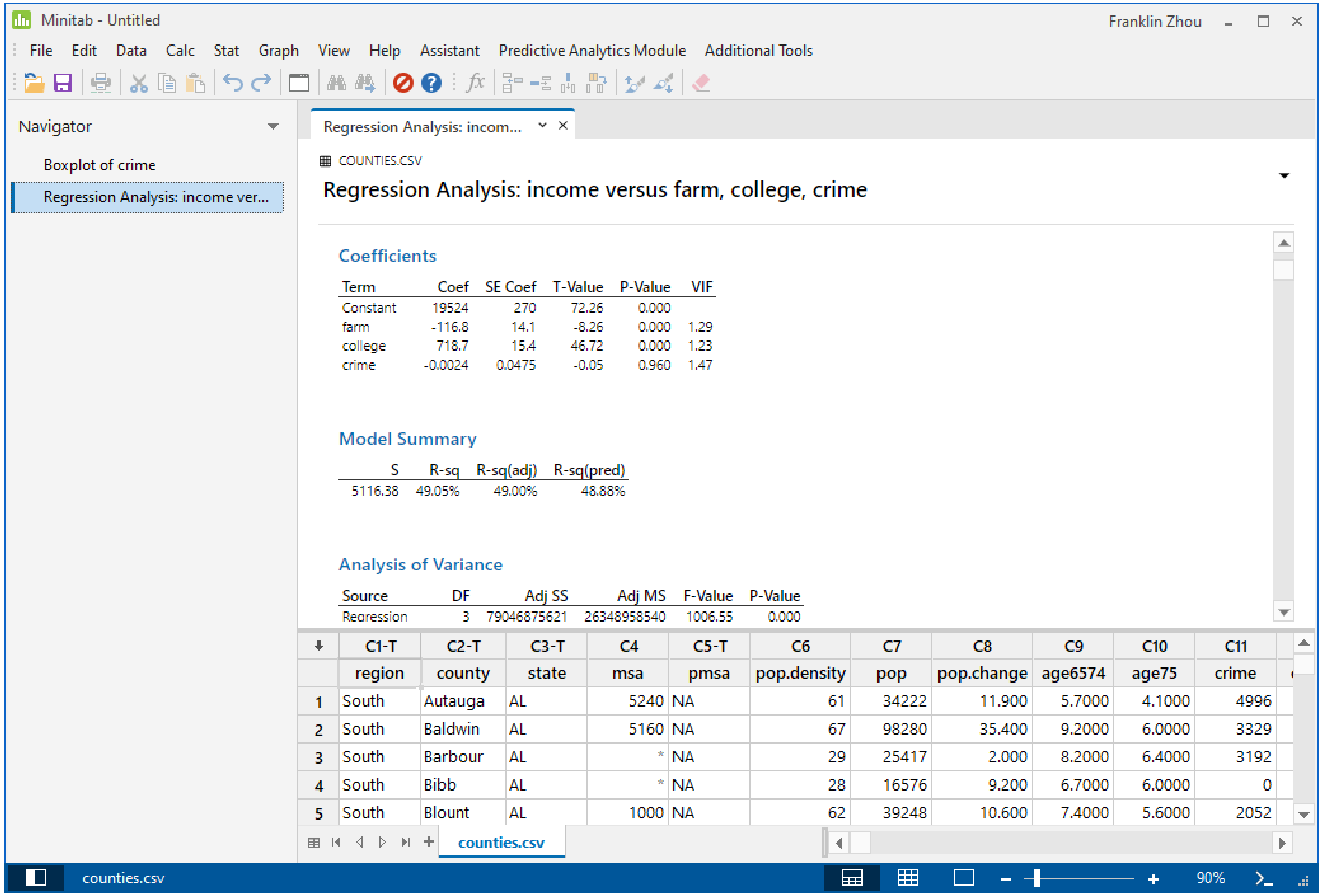
- Easy to use with GUI
- Simple and quick of building certain models, similar as JMP
- Many social science researchers (including my wife) use it.

What troubles me

- Not good at data processing
- My wife uses it.

Minitab®

- A commercial software
- With GUI
- Satisfy all basic statistical analysis needs
- Supports syntax command
- Less flexible than SPSS



Minitab - Untitled
Franklin Zhou
File Edit Data Calc Stat Graph View Help Assistant Predictive Analytics Module Additional Tools
Navigator
Boxplot of crime
Regression Analysis: income ver...
Regression Analysis: incom...
COUNTIES.CSV
counties.csv
90%

Regression Analysis: income versus farm, college, crime

Coefficients

```
Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant   19524      270    72.26    0.000
farm      -116.8     14.1    -8.26    0.000  1.29
college    718.7     15.4    46.72    0.000  1.23
crime    -0.0024   0.0475    -0.05    0.960  1.47
```

Model Summary

```
      S    R-sq  R-sq(adj)  R-sq(pred)
5116.38  49.05%     49.00%      48.88%
```

Analysis of Variance

```
Source      DF       Adj SS       Adj MS  F-Value  P-Value
Regression   3  79046875621  26348958540  1006.55    0.000
```

	C1-T	C2-T	C3-T	C4	C5-T	C6	C7	C8	C9	C10	C11
	region	county	state	msa	pmsa	pop.density	pop	pop.change	age6574	age75	crime
1	South	Autauga	AL	5240	NA	61	34222	11.900	5.7000	4.1000	4996
2	South	Baldwin	AL	5160	NA	67	98280	35.400	9.2000	6.0000	3329
3	South	Barbour	AL	*	NA	29	25417	2.000	8.2000	6.4000	3192
4	South	Bibb	AL	*	NA	28	16576	9.200	6.7000	6.0000	0
5	South	Blount	AL	1000	NA	62	39248	10.600	7.4000	5.6000	2052

Minitab®

What appeals to me

- Easy to use with GUI
- Simple and quick of building certain models
- Copy & paste the result to Word document

What troubles me

- Not good at data processing
- Does not support some advanced models (extra features costs a lot!)
- Only works within organization’s network (VPN) and one device per account.

Software	Platform	GUI	Easy to learn and use	Lots of users	Free	Programming	Best for
Excel	Windows Mac OS Web	Yes	Yes	Yes	No	No	• Data view • Pivot table • Basic statistical analysis
SAS	Windows Web*	No	No	Yes	No	Yes	• Data processing • Basic and Advanced statistical analysis • Predictive modelling
R/Python	Windows Mac OS Web*	No	No	Yes	Yes	Yes	• Data processing • Basic and Advanced statistical analysis • Predictive modelling
JMP	Windows Mac OS	Yes	Yes	No	No	No	• Advanced statistical analysis • Predictive modelling • Graphics
SPSS	Windows Mac OS	Yes	Yes	Yes	No	Some	• Basic and Advanced statistical analysis • Graphics • Predictive modelling
Minitab	Windows	Yes	Yes	No	No	Some	• Basic statistical analysis • Graphics • Basic predictive modelling

Other tools that help with data analysis

Data visualization tools can help us to find the patterns or trends.

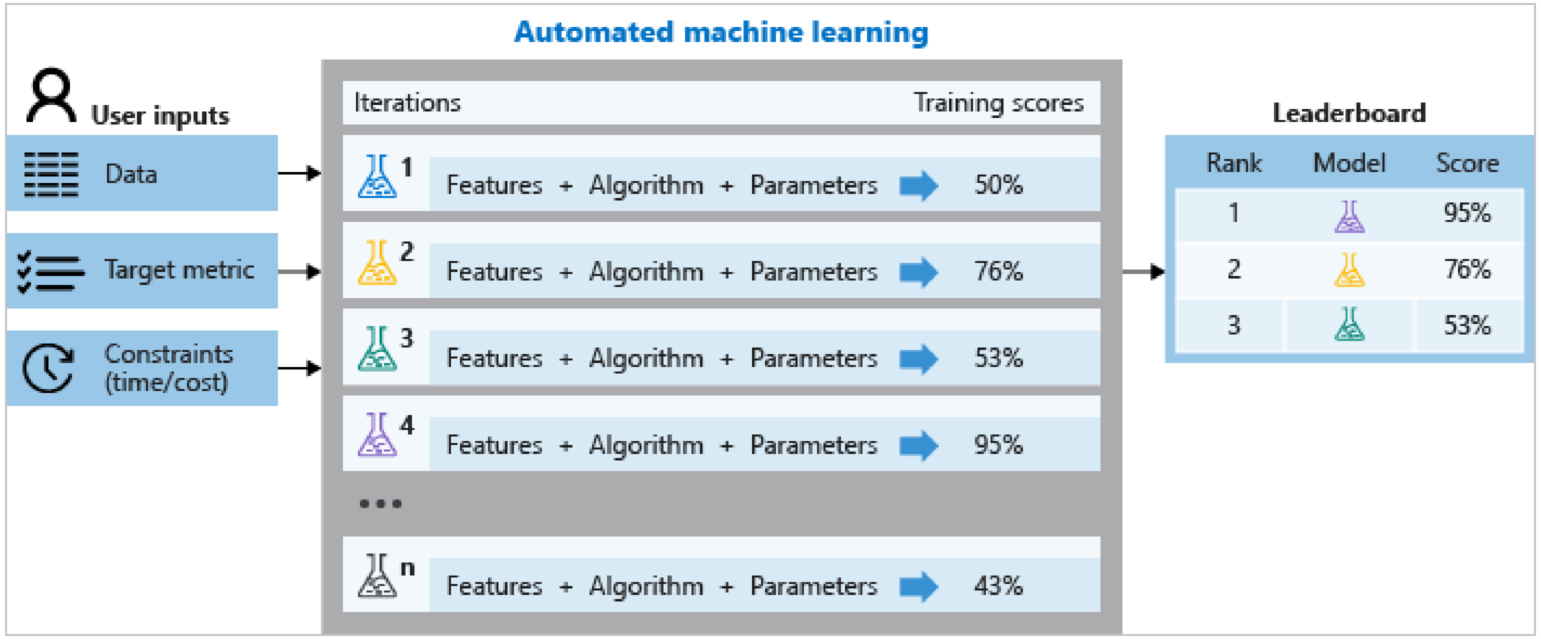
Power BI
Premium

New features of Power BI Premium

Advanced AI capabilities:

- Automated machine learning (AutoML)
  - Binary Prediction
  - Classification
  - Regression Models
  - Time-series forecasting

- Cognitive Services
  - Sentiment Analysis
  - Key Phrase Extraction
  - Language Detection
  - Image Tagging



Automated machine learning

User inputs

Data

Target metric

Constraints (time/cost)

Iterations — Training scores

1 Features + Algorithm + Parameters → 50%

2 Features + Algorithm + Parameters → 76%

3 Features + Algorithm + Parameters → 53%

4 Features + Algorithm + Parameters → 95%

...

n Features + Algorithm + Parameters → 43%

Leaderboard

Rank	Model	Score
1	⚗	95%
2	⚗	76%
3	⚗	53%

Share your experience and opinion

Thank you

ECU®

Thank you for attending the 2024 NCAIR Annual Conference!

There's a QR code in your program for a conference evaluation form. You'll also get an e-mail following the conference with a link to the form, which will be available until 4/30.

At your earliest convenience, please take this opportunity to let the planning committee know your thoughts about this year's conference and where you would like to meet next year.