You will spend nearly 70% of your time doing this!

Organize & pre-process your Data

Rutgers Libraries - NB Data Science Workshop Series

Pratiksha Sharma

Oct 13, 2022

Fall 2022 Hours

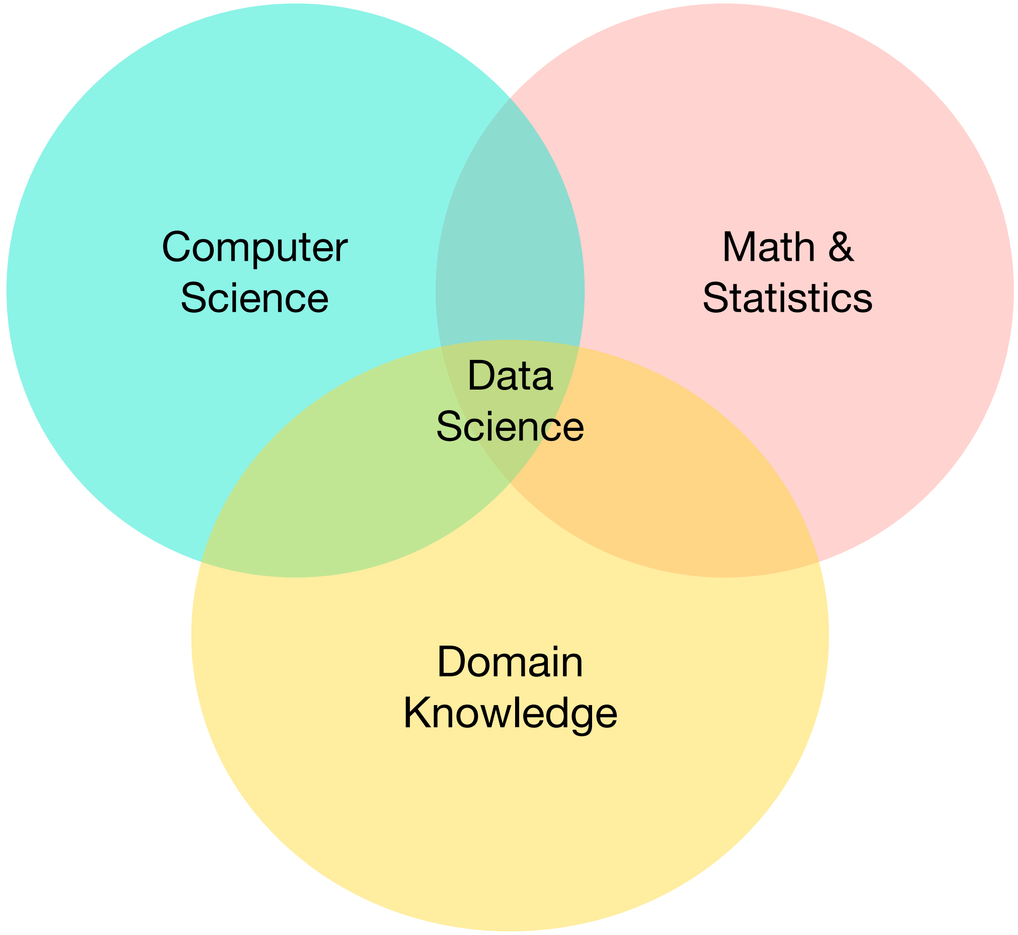
Pratiksha Sharma - Data Science Graduate Specialist

Email: pratiksha.sharma@rutgers.edu
Topics: Data Science, Tableau, Python, SQL & NoSQL Databases
Office Hours (by appointment):
Thursday 12:30 - 01:00 pm (on days when workshop ends at 12:30 pm)
Thursday 01:00 - 01:30 pm (on days when workshop ends at 01:00 pm)
General Consultation: Request an appointment via email
Location:
Zoom Meeting Link
Meeting ID: 926 5210 0393
Passcode: 772895

Quick Look: Data Science

What's all the fuss about?

- A combination of Maths & Statistics, Computer Science and Domain Knowledge.
- This workshop is a combination of **Computer Science and Maths & Statistics!**
- You don't need to be an expert - but data science is a part of everyday life!



Motivation

Source: Bill Waterson | Universal Press Syndicate

Before we begin..

- We will be using MS Excel in this workshop
- Particularly some examples from the Analysis ToolPak: Load Analysis ToolPak for Excel
- You are encouraged to participate & follow along!
- Data would be provided in the Zoom chat & will be available later on with the workshop materials on: Rutgers Libguides Data Science Workshops

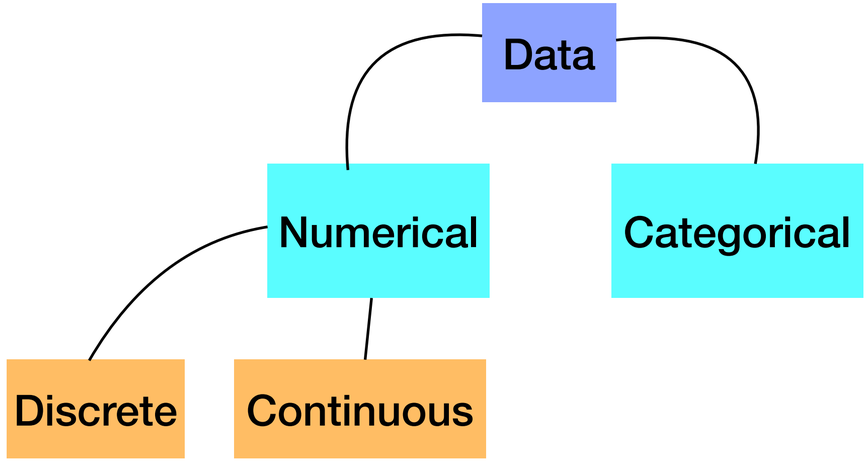
Data Cleaning, organizing & pre-processing

We will cover..

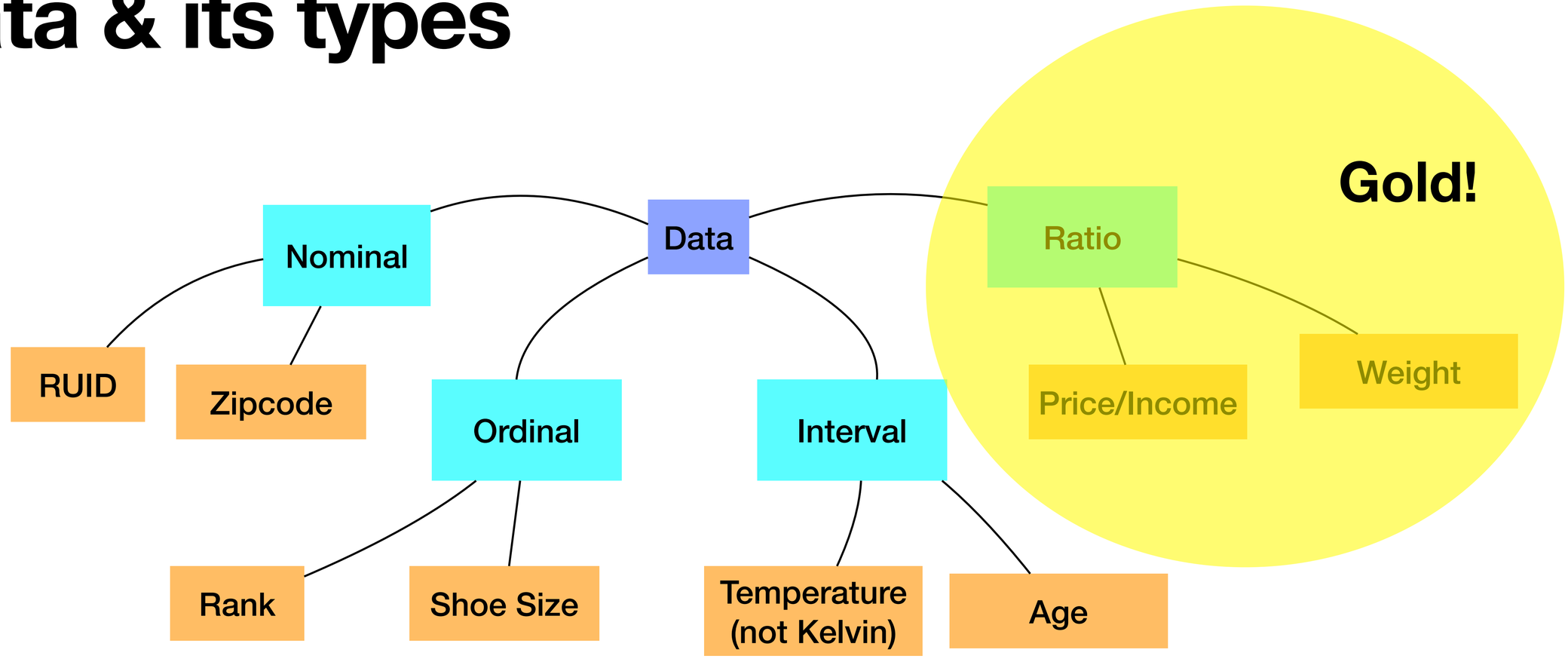
- Quick Look: Data & its types
- Current challenges
- Specific Examples: Problems & Solutions
- Linear Regression: Example with MS Excel Analysis ToolPak

Data & its types

- Data can be categorized into two types:



Data & its types



Current Challenges

- Cross-country survey data:

Customer ID	Name	Region	Phone
100	Jim Pembry	North America	732-790-6574
200	Gavin Andrews	Australia	61(460)882410
300	Shruti Gupta	India	8765607003
400	Van Dyke	UK	(+)442079357865
500	Heon Wang	China	8.62165E+11

- Do you see major problems?

Current Challenges

- The data that is generated today, has 3 main properties:

  1. Variety:

     Data is being generated and collected from various sources & various formats

  2. Volume

     Data is being generated in huge quantities; to the extent that there is a whole field of research & study dedicated to storage & retrieval of this massive amount of data

  3. Velocity

     Data is not just being generated in huge quantities, but at an unprecedented rate; to the extent that there is a whole field of research & study dedicated to high velocity collection of data

Current Challenges

- Untreated/Raw data is often disintegrated, misleading & in most cases, not useful
- Data of this kind cannot be necessarily modeled; think about creating wrong models that work for wrong data
- Unprocessed data may restrict you from performing the necessary analysis for achieving your goals: think about visualizing categorical data, outliers skewing a distribution plot, etc.
- With the data that is being generated today, manual treating may not help
- Sophisticated methods are required to deal with bigger problems

Specific Examples: Problems & Solutions

Faulty values

- The dataset on the right has faulty values
- Causes: Incorrect recording device, human error etc.
- Possible Solutions:
    1. Delete the outliers/faulty rows
    2. Substitution
        1. Mean - when the data is usually similar with no strong low or high values
        2. Median - when data has peculiar low or high values that have a potential of providing a misleading mean
        3. Mode - think about categorical data, frequency data

Student heights
158.48
143.1
155.92
155.57
147.02
271
155.37
148.04
150.98
162.82
147.59
155.75
151.13
146.93
164.79
157.14
158.84
60.46
145.06
153.26
154.06
152.37
158.12
158.11
281.5
151.77
147.96
162.05
162.91
150.2

Specific Examples: Problems & Solutions

Integrating different variety

- Come back to this example. How can you help?

Customer ID	Name	Region	Phone
100	Jim Pembry	North America	732-790-6574
200	Gavin Andrews	Australia	61(460)882410
300	Shruti Gupta	India	8765607003
400	Van Dyke	UK	(+)442079357865
500	Heon Wang	China	8.62165E+11

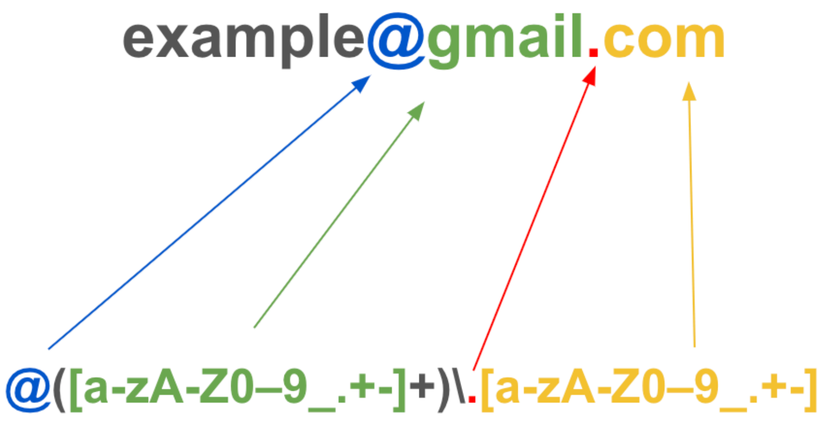
- Use the region to pick a pattern for text matching, capture important information & retain in desirable format

- The solution: Regex

Specific Examples: Problems & Solutions

Integrating different variety

- Regex stands for **Regular Expression**
- It is a ‘pattern matching’ technique
- You have data in various ‘exact’ forms but in some similar sort of fashion
- One can use little patterns to provide a general format of how data could look like
- Regex is much more powerful than this!
- Let’s jump to a movies dataset -



Specific Examples: Problems & Solutions

Look at a particular dataset

- Movies Dataset has:
    1. Columns with a lot of NA values
    2. Rows with irrelevant data
    3. Disintegrated formats of data capture
    4. Categorical to numeric conversion: Dummy conversion, one-hot encoding

Specific Examples: Problems & Solutions

Look at a particular dataset

- One-hot encoding:

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

GenreVector	W
(1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	N
(0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	N
(1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	N
(0, 0, 1, 0, 0,	

Specific Examples: Problems & Solutions

Look at a particular dataset

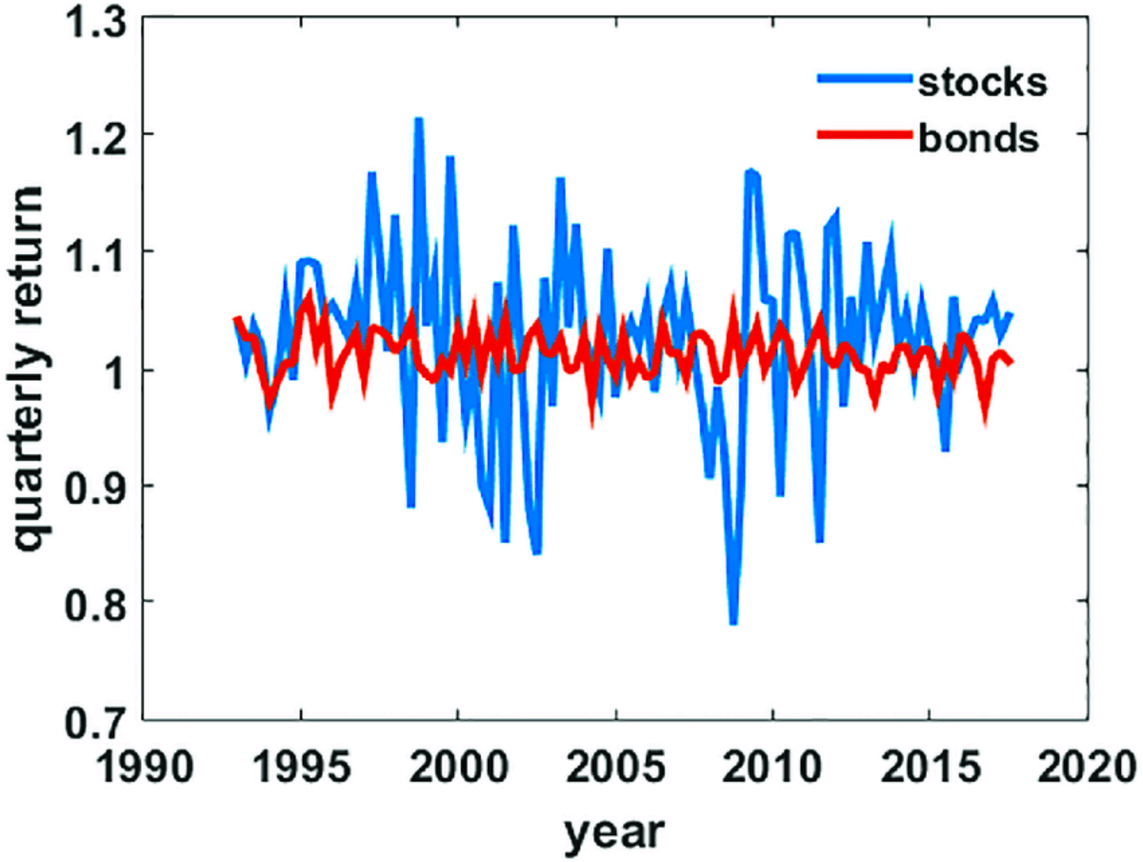
- Regex matching for more complex data
- Thinking about using a combination of different methods: first regex, then frequency filling for NA values etc.
- Think about a case where the data itself is correct; but problematic
- Normalisation
- SMOTE

Specific Examples: Problems & Solutions

Look at a particular dataset

- Normalisation:

Image Source: Guy Metcalfe | researchgate.net



Specific Examples: Problems & Solutions

SMOTE: Problem of Imbalanced Data

- Data: 997 cat pictures, 3 dog pictures
- You make a model: (the specifics of the model are out of scope)
- You make a model with superb accuracy!
- A model that looks at every image & says - this is a CAT!
- Accuracy = 99.7 % AMAZING?!
- What about the Dog class?
- Enter: SMOTE

Specific Examples: Problems & Solutions

SMOTE: Problem of Imbalanced Data

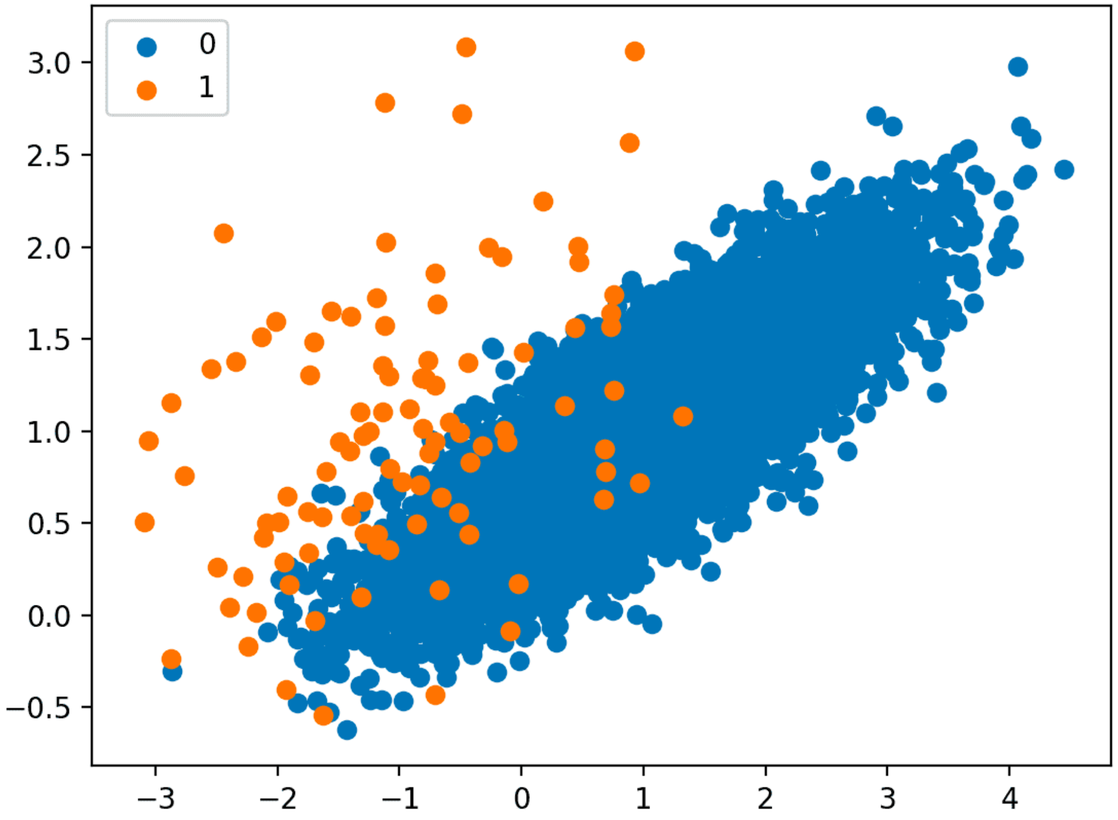


Image Source: https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification/

Specific Examples: Problems & Solutions

Problem of Imbalanced Data

- SMOTE: **S**ynthetic **M**inority **O**versampling **Te**chnique
- For imbalanced classification cases, SMOTE has the goal to increase more examples of underrepresented data i.e. minority data
- Some applications: Fraud Classification, Rare Virus finding etc.
- SMOTE uses a technique where it will look for data points most similar to a data point from a minority class, and will generate synthetic samples between the actual and the neighboring data points
- Repeat SMOTE to say double, or triple current representation of minority data to make a better model

Specific Examples: Problems & Solutions

Feature Engineering

- Sometimes, you have data that by itself does not make a lot of sense or does not necessarily give you a good model
- Transformations don’t improve performance too
- What do you do?
- Feature engineering concentrates on combining one or more multiple features to create a higher, better performing feature

Specific Examples: Problems & Solutions

Feature Engineering

- Sometimes, you have data that by itself does not make a lot of sense or does not necessarily give you a good model
- Transformations don’t improve performance too
- What do you do?
- Feature engineering concentrates on combining one or more multiple features to create a higher, better performing feature

Linear Regression

- Some feauture engineering
- Discussion of dealing with categorical data
- Linear Regression model with Numeric data

Takeway:

Modeling is not as straightforward as it seems!

Upcoming Workshops

https://libcal.rutgers.edu/nblworkshops

- The Power of Visual Storytelling: Learning Tableau Public : Oct 20
- “Make your computer work for you!" Learn how to use Python to program your tasks- Part 1 : Oct 27
- “Make your computer work for you!" Explore popular Data Science libraries in Python - Part 2 : Nov 03

Feedback Form

https://rutgers.libwizard.com/f/graduate_specialist_feedback