https://tinyurl.com/CADS-Intro-To-DataScience

Intro to Data Science with an Example in Python

Carolina Analytics & Data Science

https://tinyurl.com/CADS-Intro-To-DataScience

WELCOME!

Attendance

https://tinyurl.com/CADS-Intro-To-DS

https://tinyurl.com/CADS-Intro-To-DataScience

Warm Up

https://tinyurl.com/CADS-Intro-To-DataScience

Survey

- How experienced are you in data science?
- Respond on PollEv.com/zwei596

What level are you currently at?

Had plenty of experiences on data science	**A**
Self-studied data science before	**B**
Took classes on statistics / data science before	**C**
Beginner but interested in data science	**D**

https://tinyurl.com/CADS-Intro-To-DataScience

Share Your Stories

- Share with us your experiences about data science , your comprehension of data science, what do you expect from this workshop , or anything!
- Respond on PollEv.com/zwei596 or in chats.

https://tinyurl.com/CADS-Intro-To-DataScience

What does a data scientist do?

https://tinyurl.com/CADS-Intro-To-DataScience

- Use scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data.

- Skills:
  - Programming skills
  - Statistics (probability, machine learning, etc.)
  - Communication skills
  - Persistence (80% of a data scientist's valuable time is spent simply finding, cleaning, and organizing data)

https://tinyurl.com/CADS-Intro-To-DataScience

Why are data scientists important?

https://tinyurl.com/CADS-Intro-To-DataScience

The Sexiest Job of the 21st Century

- Wide applications in …
  - Healthcare (early diagnosis, accelerate development of medicine, …)
  - Sports (Sports analytics, sports betting firms, professional sports teams)
  - Search engines, targeted advertisements, facial recognition, self-driving cars, arts (Google Quick Draw, Google Magenta Project)
  - ...
- Increasing demand
  - 2017 report by IBM: predicted that the number of analytics and data science positions in the U.S. alone would increase by 364,000, to 2,729,000 by 2020
  - 2019 rankings by Linkedin: “data scientist” the No. 1 most promising job in the U.S; a 56% rise in job openings for data scientists
- Quickly evolving

https://tinyurl.com/CADS-Intro-To-DataScience

How to take the first step?

https://tinyurl.com/CADS-Intro-To-DataScience

- Courses and textbooks:
  - Courses offered by UNC
  - Online resources (statistics, python, R, SQL -> Tableau -> machine learning...)
    - Coursera (Andrew's Machine Learning class)
    - W3schools
  - Textbook
    - Introduction to Statistics & Data Analysis
    - Introductory Econometrics
- Datasets for practice:
  - AWS
  - Economic data: CEIC, IMF
  - A categorized and organized list of data sources
- Research:
  - Find a topic that you are interested in!
  - Use your contacts
  - Resources at UNC
  - Pioneer Academics

https://tinyurl.com/CADS-Intro-To-DataScience

Example using Python

https://tinyurl.com/CADS-Intro-To-DataScience

Where do you start?

- Ask a Question
- Find the Data
- Read in the Data
- Clean the Data
- Perform you Analysis
- Visualize your Findings

https://tinyurl.com/CADS-Intro-To-DataScience

The Tools We'll Use

Libraries:

- Pandas:
  - Data manipulation and Analysis
- Matplotlib:
  - Plotting library

Google Colab:

- Write and execute Python code through your browser
- Has many libraries for Data Science built in
  - Including Pandas & Matplotlib

Github:

- You can find the code and data set at this repository!
  - One file is completely filled in
  - The other has blanks for you to fill in

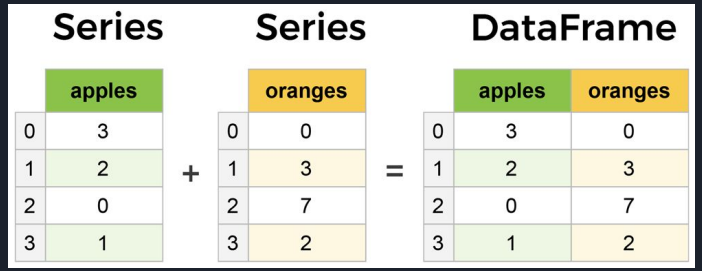
https://tinyurl.com/CADS-Intro-To-DataScience

Our Question:
What undergraduate fields of study pay the most at UNC-CH?

https://tinyurl.com/CADS-Intro-To-DataScience

The Data

- From the Department of Education
- Data on fields of study at US college institutions
- CSV format - comma separated values
  - Easy to use for data science
- Data Frames:
  - 2 dimensional data structure
  - Aligned in rows and columns



	Series		Series		DataFrame	
	apples		oranges		apples	oranges
0	3	+	0	=	3	0
1	2		3		2	3
2	0		7		0	7
3	1		2		1	2

https://tinyurl.com/CADS-Intro-To-DataScience

Pandas Basics

- **read_csv():**
  - Reads a csv file into a dataframe
- **DataFrame.head():**
  - Returns the first 5 rows of a data frame
- **DataFrame[‘Column Name’]**
  - This is how you select a column from a data frame
- **DataFrame.loc[]:**
  - Access a group of rows & columns by a label
- **DataFrame.iloc[]:**
  - Access a group of rows & columns by integer location based indexing
- **DataFrame.sort_values():**
  - Sorts rows in ascending or descending order based on the value of a certain column

https://tinyurl.com/CADS-Intro-To-DataScience

Matplotlib Basics

- **bar():**
  - Creates a bar chart when you input an x-axis and a height
- **scatter():**
  - Creates a scatter plot when you input x and y variables
- **xticks() or yticks():**
  - Allows you to manipulate the formatting of the ticks on the x or y axis
- **xlabel() or ylabel():**
  - Allows you to add and format an axis label
- **title():**
  - Allows you to add and format a title for the plot

https://tinyurl.com/CADS-Intro-To-DataScience

Upcoming Workshops and Events

- Research Panel Questions (Submit Questions)
- Nov. 6th - Course Registration Workshop (Submit Questions)
- Nov. 11th - Semester Recap & End of Fall Semester “Party”

https://tinyurl.com/CADS-Intro-To-DataScience

Stay Connected!

Join our Slack

Join as a member on HeelLife

Sign up on our Listserv

Follow us on LinkedIn

Like us on Facebook

Follow us on Instagram

Visit our Website: http://carolinadata.unc.edu/

https://tinyurl.com/CADS-Intro-To-DataScience

Thank you!