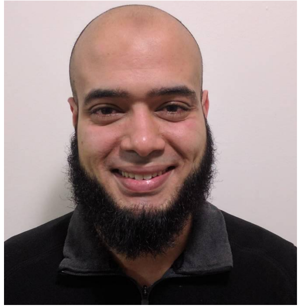
Machine Learning

Pandas Library

Mostafa S. Ibrahim
Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher
PhD from Simon Fraser University - Canada
Bachelor / MSc from Cairo University - Egypt
Ex-(Software Engineer / ICPC World Finalist)



Pandas

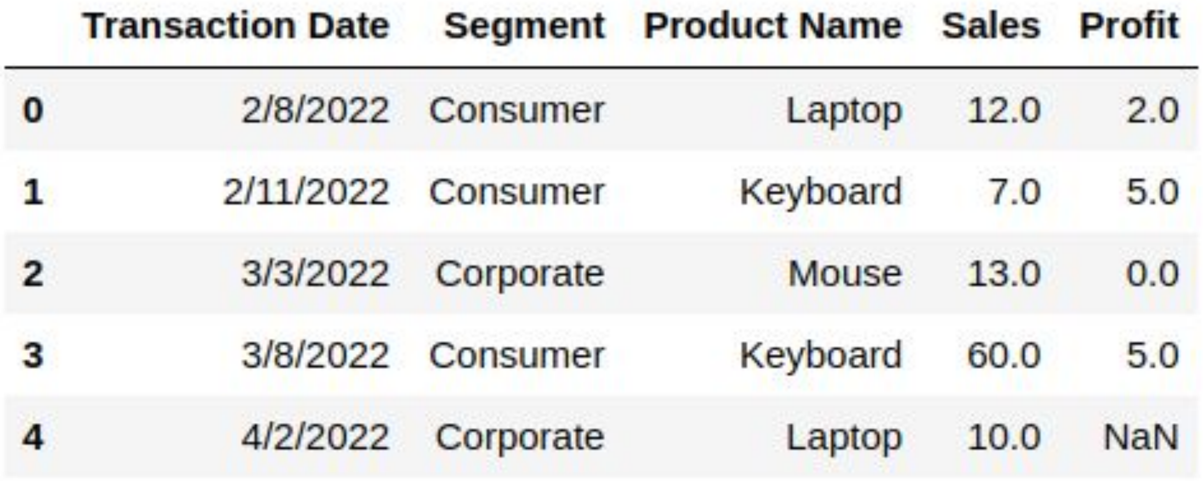
- Pandas is package built on top of Numpy and used in processing datasets
    - pip install pandas
- Pandas vs Numpy: critical difference
    - Panda is **column major**-order data: efficient to access using columns NOT rows
    - Numpy is **row major**-order data: efficient to access using rows NOT columns
- In general, useful tool to process cvs/tabular files
- It has intensive library functions that do a lot for tabular data
- A common tool for investigating data for data scientist/analyst

Pandas and your job

- CV and NLP fields use it in a limited way
    - E.g. save/load performance metrics in a file
- Other ML folks, use it to explore the features: Basic usage
    - Are there missing values? Columns stats
- Data analytics / data science
    - Can run intensive analysis in the data to find insights

Tabular Data

- 2D table: columns represents features and rows represents examples
  - Examples can have problems: e.g. duplicate, missing values, etc
- We need to have some understanding for the data first
  - Or deep understanding for data science insights



	Transaction Date	Segment	Product Name	Sales	Profit
0	2/8/2022	Consumer	Laptop	12.0	2.0
1	2/11/2022	Consumer	Keyboard	7.0	5.0
2	3/3/2022	Corporate	Mouse	13.0	0.0
3	3/8/2022	Consumer	Keyboard	60.0	5.0
4	4/2/2022	Corporate	Laptop	10.0	NaN

Basic Data Understanding

- Given a tabular data, you may understand it in 2 steps
- 1) Individually, column by column
- 2) Pair/Group of column understanding

Per Column

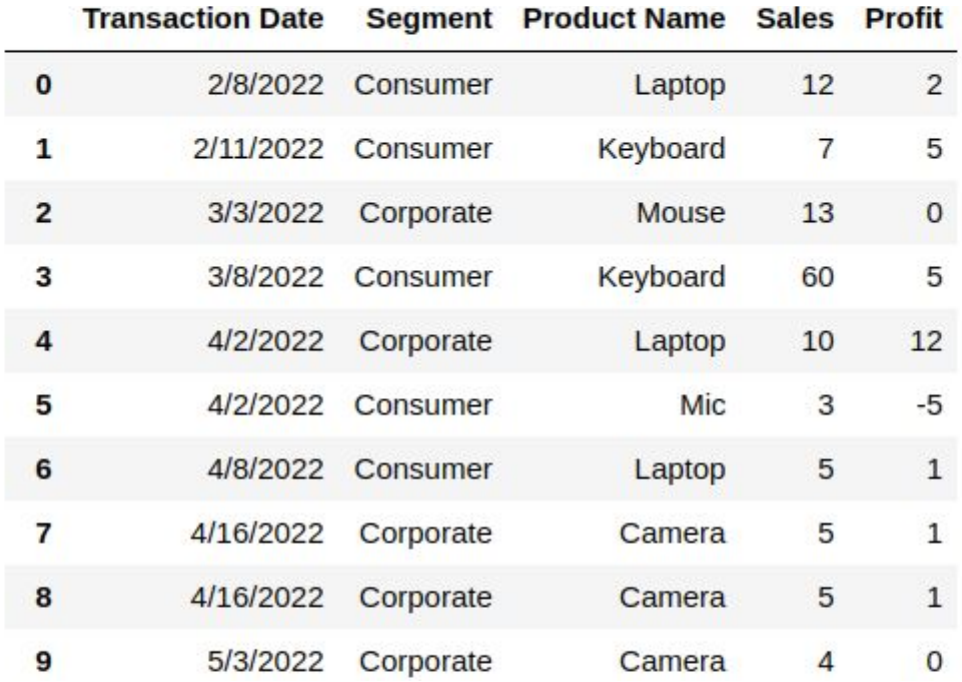
- Investigate the values of a column
    - Some columns are numeric
    - Others can be: date or categories
- How many null values?
- How many duplicate values? Unique values?
- What are the basic statistics: mean, std, quartiles, etc
- Raw filtration
    - You might want also to filter rows
    - For example, keep only rows that has > 1000 in revenue
    - Or filter sales of only top 75% percentile (or any other quantity)
- (Later) For classification problems, check the imbalance of the target label

Group of Columns

- Think about **questions** that can help you extract information
- Determine the **columns** of interest
- Determine one or more **major** columns among them
- Group the data by these major columns and **aggregate** specific features
    - For example, group the data by customer-segment and see the mean/max revenue
        - You may sort **ascending or descending** to print elements (e.g. find the top products)
    - **Visualize** your findings
- **Temporal** information
    - If you have **date** (e.g. order date, transaction date), what is the progress of the columns of interest against the date
- You may compute/visualize the correlation matrix of numeric values

Question

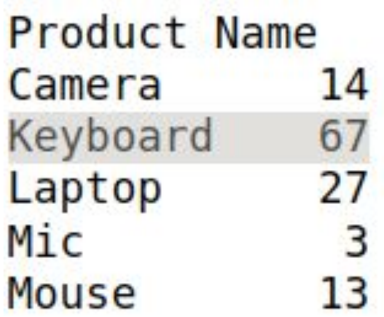
- Which product has the largest sales?
  - Which **columns** do we need?



	Transaction Date	Segment	Product Name	Sales	Profit
0	2/8/2022	Consumer	Laptop	12	2
1	2/11/2022	Consumer	Keyboard	7	5
2	3/3/2022	Corporate	Mouse	13	0
3	3/8/2022	Consumer	Keyboard	60	5
4	4/2/2022	Corporate	Laptop	10	12
5	4/2/2022	Consumer	Mic	3	-5
6	4/8/2022	Consumer	Laptop	5	1
7	4/16/2022	Corporate	Camera	5	1
8	4/16/2022	Corporate	Camera	5	1
9	5/3/2022	Corporate	Camera	4	0

Answer

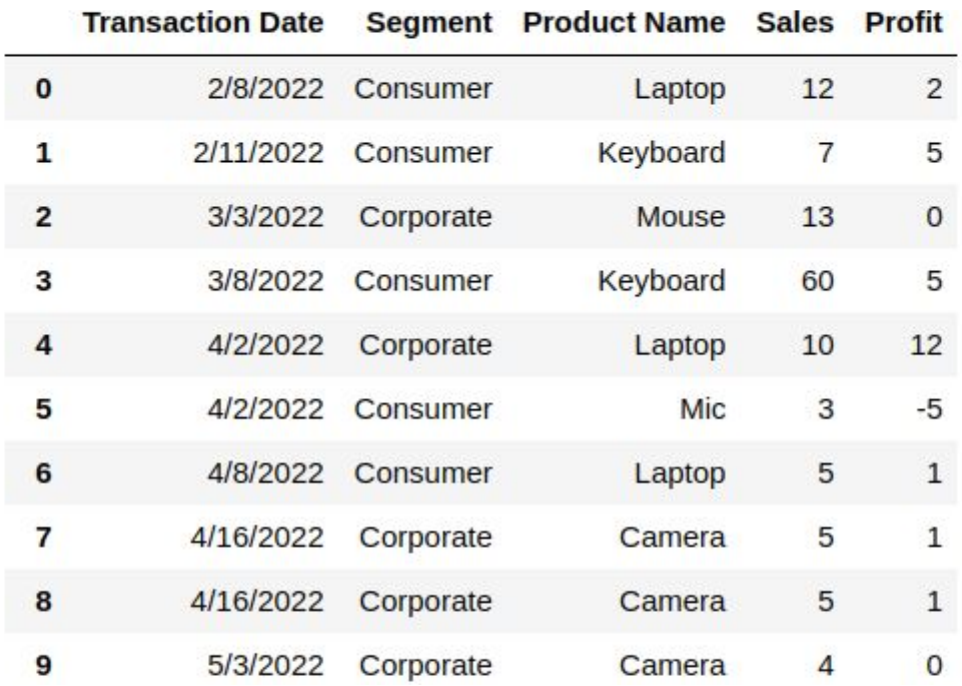
- First, we need to **group** the data by the Product name
    - Camera: rows [7, 8, 9]
    - Keyboard rows: [1, 3]
    - Laptop rows [0, 4, 6]
    - Mic rows [5]
    - Mouse: rows [2]
- Second, sum the sales of each group and find the largest one!



```
Product Name
Camera      14
Keyboard    67
Laptop      27
Mic          3
Mouse       13
```

Question

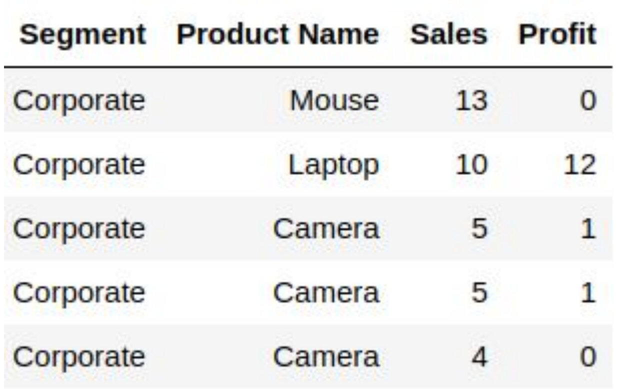
- In the Corporate segment, which product has the largest profit?
    - Which **columns** do we need?



	Transaction Date	Segment	Product Name	Sales	Profit
0	2/8/2022	Consumer	Laptop	12	2
1	2/11/2022	Consumer	Keyboard	7	5
2	3/3/2022	Corporate	Mouse	13	0
3	3/8/2022	Consumer	Keyboard	60	5
4	4/2/2022	Corporate	Laptop	10	12
5	4/2/2022	Consumer	Mic	3	-5
6	4/8/2022	Consumer	Laptop	5	1
7	4/16/2022	Corporate	Camera	5	1
8	4/16/2022	Corporate	Camera	5	1
9	5/3/2022	Corporate	Camera	4	0

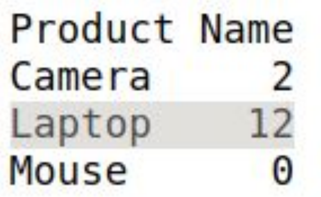
Answer

- First, filter all the rows to select only Segment = Corporate



Segment	Product Name	Sales	Profit
Corporate	Mouse	13	0
Corporate	Laptop	10	12
Corporate	Camera	5	1
Corporate	Camera	5	1
Corporate	Camera	4	0

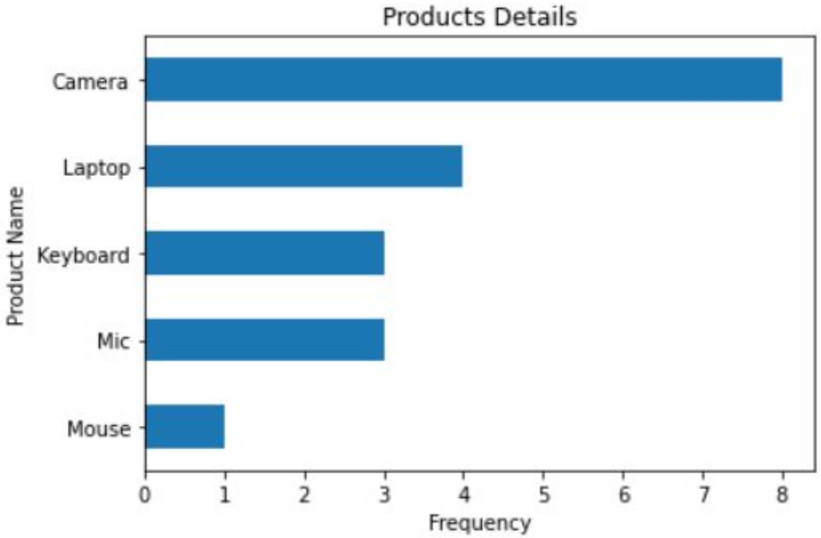
- Now, group by the product and sum the profit for each group

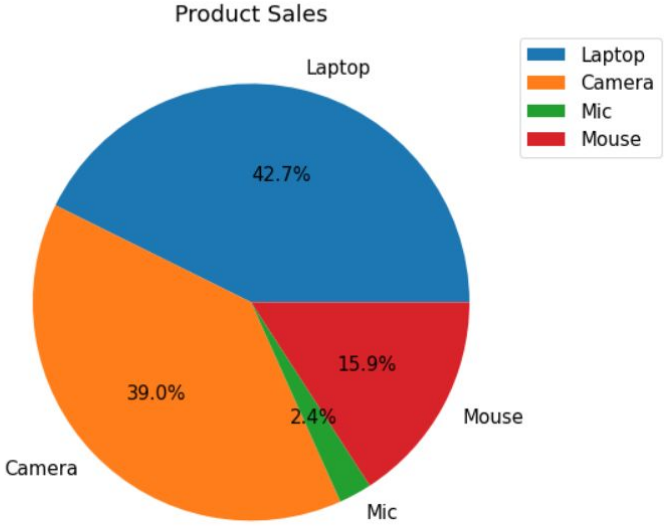


```
Product Name
Camera       2
Laptop      12
Mouse        0
```

Visualization

- For the different information, generate suitable visualizations
  - In Data science, you need to understand and think about these visualizations





Relevant Materials

- Written: link
- Wanna learn some function, google it: e.g. pandas set_index
  - https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.set_index.html
- Don't keep learning. Use docs / see examples when need
  - Get some basics.
  - Get overview of what can be done
  - Google/search later
- How to Read a Correlation Matrix
- **For Data Analysts / Data Science**
  - Coursera: Google Data Analytics Professional Certificate
  - Book: head first data Analysis / video

Jupyter Tour

“Acquire knowledge and impart it to the people.”

“Seek knowledge from the Cradle to the Grave.”