

## Business Intelligence Techniques and Applications

### Session 1b. Python and Data Analysis Basics

Renyu (Philip) Zhang

1

## Python and Jupyter

- Python: Very close to English, so not hard to learn.
- Most widely used programming language by data scientists.
  - A huge community with extensive external packages (especially ML & AI).
  - Very easy to find solutions and support when running into problems.
- Jupyter Notebook.
  - A web-based interactive computing environment.
  - Use Anaconda to install Jupyter.
  - Install Python: <https://www.python.org/downloads/>
  - Install Anaconda: <https://docs.anaconda.com/anaconda/install/>
  - Alternative: Google Colab  
<https://colab.research.google.com/notebooks/intro.ipynb>
- Wisely use **Google** whenever you face any unexpected issues.



2

2

## Code Distribution via GitHub

- GitHub (<https://github.com/>) is a platform for storing code and conducting version control for software development.
  - We use the GitHub for code distribution: <https://github.com/DSME6756/BA-W2021>
  - Please join the GitHub Classroom by clicking: <https://classroom.github.com/a/RCI4-aC>
  - Read this doc for getting started with GitHub: <https://docs.github.com/en/get-started/quickstart>
- Our GitHub site has only 1 repository (a.k.a. mono-repo).
  - Please remember to **pull from this repo to your own computer everyday**.



3

3

## Data Base and SQL

- In practice, data are stored in the data base(s) of the firm/organization you work for. You need to pull them out using SQL (Structured Query Language).
  - SQL is the standard language for storing, manipulating and retrieving data in databases.
  - An important technique to gain edge on the job market.
- SQL tutorial: <https://www.w3schools.com/sql/>, <https://www.sqltutorial.org/>, <http://www.mathcs.emory.edu/~cheung/Courses/377/Others/tutorial.pdf>, <https://cs.uwaterloo.ca/~tozsu/courses/CS338/lectures/4%20Basic%20SQL.pdf>
- SQL questions (and reference solutions) from Leetcode: <https://github.com/DSME6756/BA-W2021/tree/main/SQL%20References/Leetcode>



4

4

## NumPy and Pandas

---

- Basic libraries in Python to work with arrays (i.e., matrices).
- Foundations of data analysis.
  - Most data analytics tools are based on NumPy and Pandas.
  - <https://numpy.org/>
  - <https://pandas.pydata.org/pandas-docs/stable/index.html>
- Important operations:
  - Data loading, extracting, joining, aggregating, etc.
- In this course, you should be reasonably familiar with them.
- Reference book: *Python for Data Analysis*, 2<sup>nd</sup> Edition, by Wes McKinney (available on GitHub).

5

5

## Homework

---

- Problem Set 1, due at 9:30AM, December 16, Thursday
  - Submit the solutions with code in a **Jupyter Notebook** on **Blackboard**.
- Next week: Merging data, data visualization, and linear regression.

6

6