

Introduction to Data Science

Course 094201

Lab 2:


Basic Python and Data Analysis

Spring 2020

Welcome to PyCharm

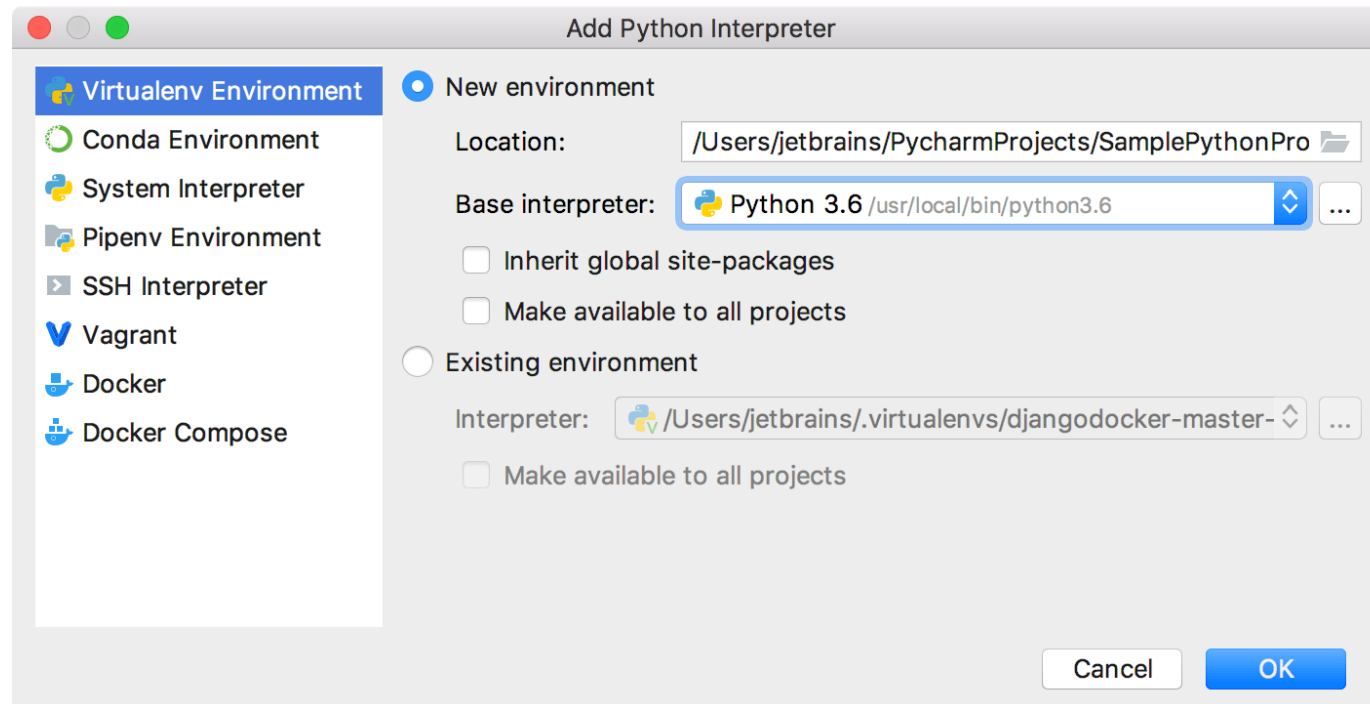
- Download lab 2 content from Moodle
- Unzip lab2_for_students.tgz
- PyCharm will be our workspace for Python this semester
- Open PyCharm (Applications->Programming->PyCharm)
- Select: “Import Project from Sources”
- Select the project you’ve just unzipped.
- Configure a virtual environment (Next Slides)
- Run the executable from the terminal (as in lab1).

Configure a virtual environment

- Ensure that you have downloaded and installed [Python](#) on your computer.
- Open the **Add Python Interpreter** dialog by either way:
 - When have an open file in the **Editor**, the most convenient way is to use the **Python Interpreter** widget in the status bar. Click the widget and select **Add Interpreter ...**
 - Open the Settings / Preferences Dialog by pressing Ctrl+Alt+S or by choosing **File | Settings** for Windows and Linux or **PyCharm | Preferences** for macOS. In the **Settings/Preferences** dialog Ctrl+Alt+S, select **Project <project name> | Project Interpreter**. Click the  icon and select Add.



Configure a virtual environment

- In the left-hand pane of the **Add Python Interpreter** dialog, select **Virtualenv Environment**. The following actions depend on whether the virtual environment existed before.




Configure a virtual environment

If **New environment** is selected:

1. Specify the location of the new virtual environment in the text field, or click  and find location in your file system. Note that the directory where the new virtual environment should be located, must be empty!
2. Choose the base interpreter from the list, or click  and find a Python executable in the your file system.
3. Select the Inherit global site-packages checkbox if you want to inherit your global site-packages directory. This checkbox corresponds to the **--system-site-packages** option of the virtualenv tool.
4. Select the **Make available to all projects** checkbox, if needed.

Configure a virtual environment

If **Existing environment** is selected:

1. Expand the Interpreter list and select any of the existing interpreters.
Alternatively, click  and specify a path to the Python executable in your file system, for example, **C:\Python36\python.exe**.
 2. Select the checkbox **Make available to all projects**, if needed.
- Click **OK** to complete the task.
 - You can create as many virtual environments as required. To easily tell them from each other, use different names.

The Data

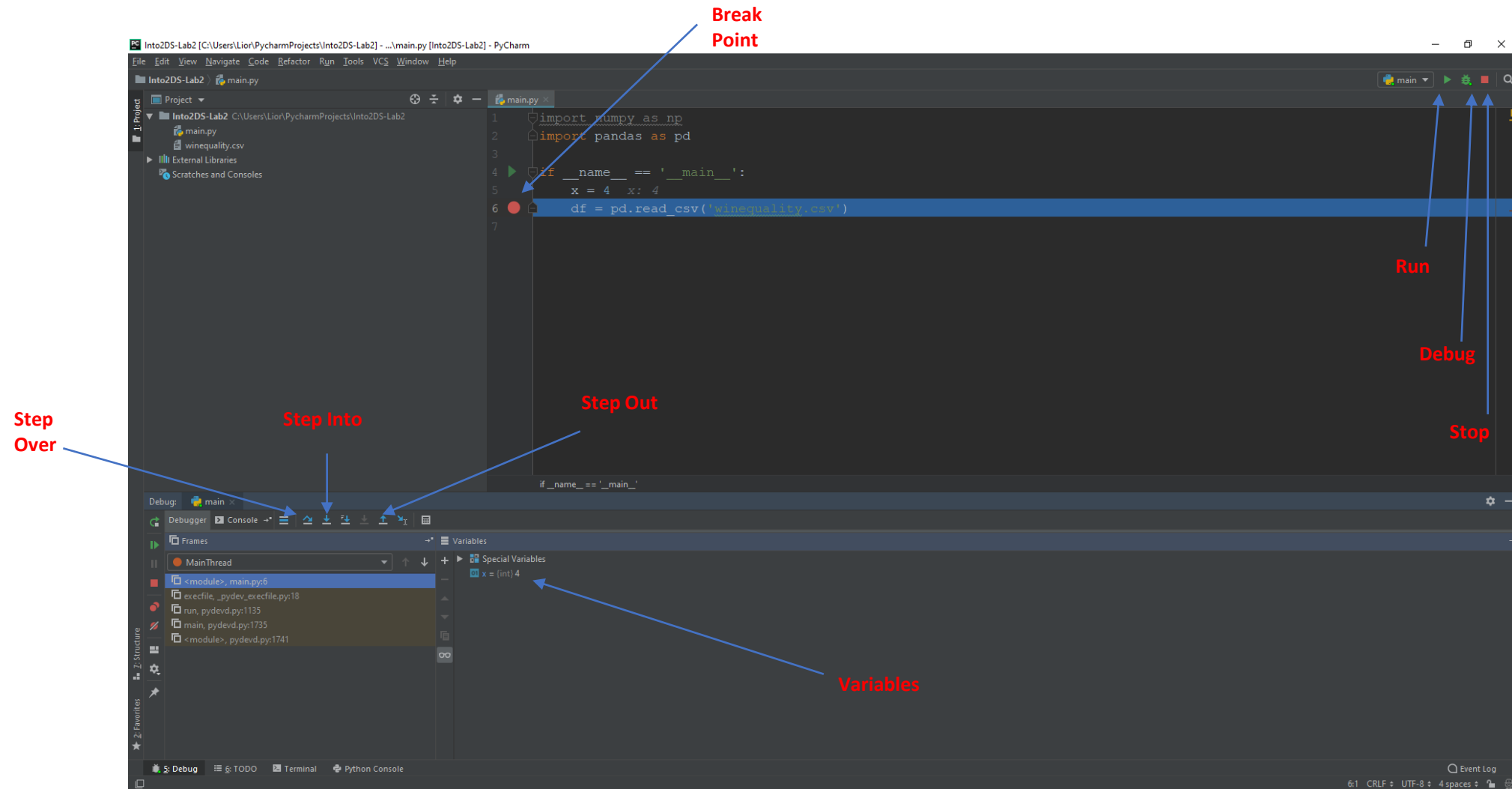
רקע:

- לרשותכם קובץ נתונים על איכות יין
- הנכם מעוניינים לבצע חקר של הקשרים בין המשתנים השונים בקובץ, כגון רמת חומציות, רמת הסוכר והאלכוהול וכו'. איכות היין (המשתנה quality) הוא מדד סובייקטיבי הניתן על סמך שיפוט אנושי.
- שימו לב שכאשר מדובר בבעיות חיזוי, סטטיסטיקאים נוטים לחלק את המשתנים (covariates\variables) למשתנה תלוי (dependent) – שזה המשתנה שמנסים לחזות ומשתנים מסבירים (explanatory). למשל, אם המטרה הייתה חיזוי איכות היין, אזי המשתנה של האיכות היה התלוי והמשתנים האחרים בקובץ היו המסבירים. לעתים נקרא המשתנה התלוי גם response variable.
- אנשי Machine Learning (זה אנחנו) משתמשים במינוח אחר בדר"כ. למשתנה תלוי קוראים label, class או outcome variable. ואילו למשתנים המסיבירים קוראים features (תכונות) או predictors. לעתים משתמשים במושגים אלו במעורב.

Understanding the Data

- לפני שמתחילים לעבוד עם דאטה, מומלץ לאסוף עליה כמה שיותר מידע.
- הסתכלות מהירה על השורות הראשונות בקובץ הנתונים או על סוג האיברים בו יכולה לספק מידע (במידה מצומצמת).
- נלמד על פקודות בסיסיות בפייתון שעושות זאת היום. ללא תלות בפייתון אפשר להשתמש בפקודות bash כמו head שלמדנו במעבדה הקודמת.
- בקובץ main.py מצויה שורה הטוענת את קובץ הנתונים.

Debugging with PyCharm



Debugging with PyCharm

- To stop the execution of the program at some point and observe the current state of our program.
- Set a few “break points” in main.cpp and debug the program
- Set a break points to observe:
 1. The value of the 4th feature in the 3rd iteration.
 2. The name of the 6th feature.

Data Analysis

- The file Statistics.py should contain some statistical measures we learned in class
- The implementation of the Median function is given. Your task is to complete the coding of the other measures

Data Analysis

- Do it yourself
- Implement the following
 - Mean: a function that receives a const reference to a vector and returns its mean.

Mean (\bar{x})

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Data Analysis

- Do it yourself
- Implement the following:
 - Variance: a function that receives a const reference to vector and returns its variance.

The sample variance:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$