Financial Econometric Analysis
Lecture 1 - Introduction

Fang-Chang Kuo

National Chung Cheng University

September 12, 2020

Outline

- Introduction
  - Syllabus
  - Business Question and Solution: A/B testing
  - Python
  - Integrated Development Environment

Syllabus

Overview of this Course

- **Overview:**
  - This is an advanced applied economics course at graduate level.
  - I will introduce machine learning tools used in **marketing**, **economics**, and **data science**.
  - We will applies these tools to analyzing various real and fake data-sets.
- **Goals:**
  - provide students with knowledge and toolkit to
    1. conduct empirical analyses that involves inference and prediction,
    2. evaluate economic/business policies,
    3. read quantitative analyses by other practitioners
  - learn at least one programming language: **Python**
- **Office Hour:**
  - Office: 331
  - Office hour: Sat. 11:00-13:00, or by appointment

Course Structure

- **Grading Policy:**
  - **60%** Assignments
  - **20%** Midterm
  - **20%** Final Project (Presentation + Paper)
- **Assignments:**
  - Please bring your laptops to class.
  - Assignments are extremely important in this course.
  - Students are expected to complete assignments during classes.
- **Midterm:**
  - We will disclose more details about midterm later.
- **Final Project:**
  - Just come up with a research agenda and to analyze the data using the tools and concepts discussed in class.
- **TA:**
  - There will be few TAs to help students with their assignments in class.

Course Materials

- **Slides:**
  - This course does not have a required textbook. Course slides will be provided as main materials.
- **Reference:**
  - J. M. Wooldridge. *Introductory Econometrics: A Modern Approach*. Cengage Learning, 7th edition, 2019
  - M. H. DeGroot and M. J. Schervish. *Probability and statistics*. Pearson Education, 4th edition, 2012
  - J. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009
  - P. Davis and E. Garcés. *Quantitative Techniques for Competition and Antitrust Analysis*. Princeton University Press, 2009

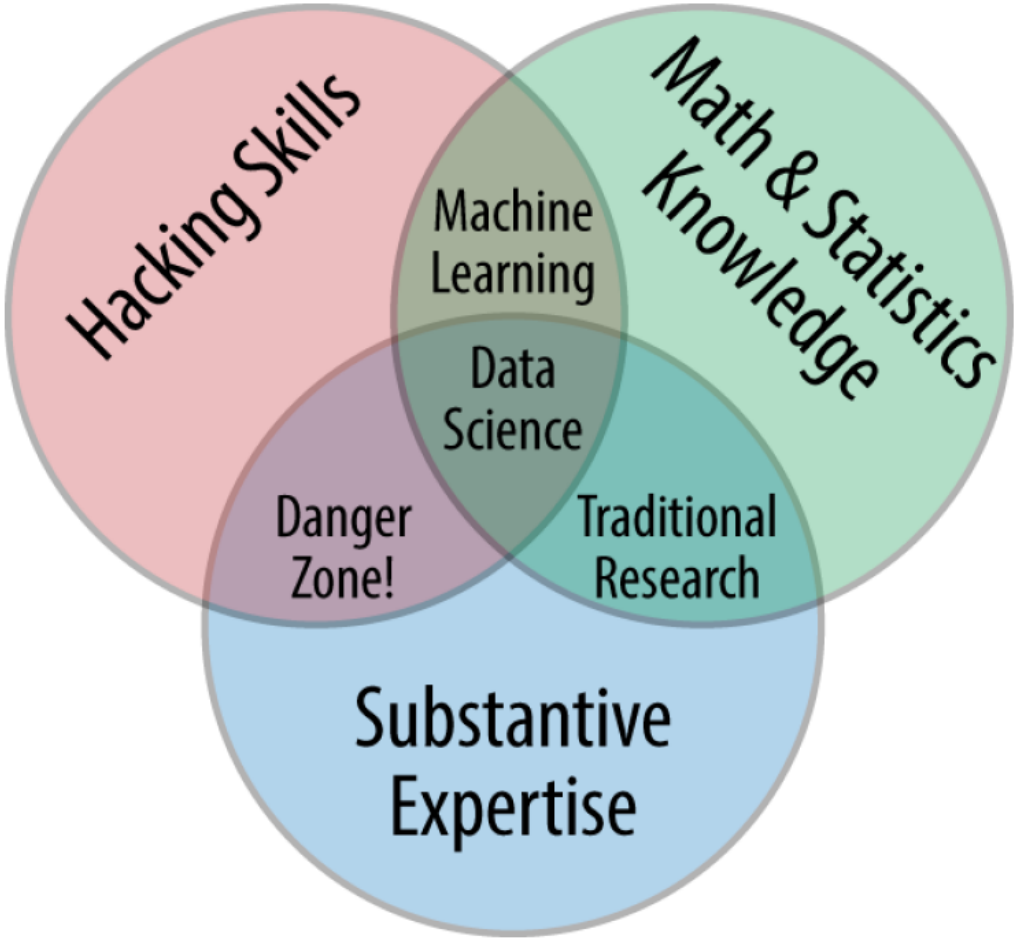
Programming Language: Python

- **Python References:**
  - W. Wesley. *Python for Data Analysis, 2nd Edition.* O'Reilly Media, Incorporated, 2017
  - J. VanderPlas. *Python Data Science Handbook: Essential Tools for Working with Data.* O'Reilly Media, Inc., 1st edition, 2016
  - J. Grus. *Data Science from Scratch: First Principles with Python.* O'Reilly Media, 2019
  - A. Müller and S. Guido. *Introduction to Machine Learning with Python: A Guide for Data Scientists.* O'Reilly Media, 2016

Course Plan

1. Introduction to Python (Numpy, Pandas, and Matplotlib)
2. Basic Data Structures, Plotting, and Visualization
3. Web Scraping (BeautifulSoup, Selenium)
4. Introduction to Machine Learning (Scikit-learn)
5. Supervised Learning: logit, KNN, Decision Tree, Random Forest, Support Vector Machine
6. Unsupervised Learning: K-means, PCA, Agglomerative Clustering

Data Science?



Business Question and Solution: A/B testing

Example: SIMCITY

- Question:
  - How to display the promotional offer in order to drive more purchases and increase revenue?
- Which one do you prefer?
  - More likely to purchase the game.
- You can have many theories or ideas of how to design the layout.
- But we only care about real sales.
- So, directly getting consumers' feedbacks is the optimal strategy.

Example: SIMCITY One Variation

Example: SIMCITY The Other Variation

A/B testing

- **A/B testing** is a way to compare two versions of something to figure out which performs better.
- The aim of A/B testing is to enable you to make incremental improvements to your website or app.
- **A/B testing** is just a fancy terminology. Essentially, it is a randomized control trial.
- The outcome variable in a A/B test is binary and you can design specific feature, or layout for your own interest.
- It is like generating a almost ideal data-set for your own use.

Example: SIMCITY Result

- Conversion Rate:
  - First Page: 10.2%
  - Second Page: 5.8%
- The difference is 43.4%.
- The promo banner actually hurt conversions.
- The results are somewhat counter-intuitive since the promo banner is meant to improve sales.
- However, it went the opposite way.
- It's good for them to run an A/B test just before full release.

Notes

- To conduct a valuable A/B test, it's crucial that you limit changes to one variable.
- There are other considerations:
- How many traffics/observations do you need?
  - It is a trade-off of cost and accuracy.
- Is the difference statistically significant?
  - We will talk about this immediately.
- Is the difference economically significant?
  - Based on your estimates on sales difference

Test of Differences in Proportion

- In statistics, A/B testing is just a two samples test in proportion/conversion.
- We can think of each variation as a Bernoulli trial where p_A is the probability of conversion for A.
- Let's say that N_A people see A, and that n_A of them click it.
- Then if N_A is large, we know that n_A/N_A is approximately a normal random variable with mean p_A.
- Similarly, for alternative B.

Test of Differences in Proportion

- Let's review some basics of hypothesis testings.
- Hypothesis: $H_0 : p_A = p_B$
- Standard deviation:

$$S_{p_A - p_B} = \sqrt{p(1-p)\left(\frac{1}{N_A} + \frac{1}{N_B}\right)}, \quad p = \frac{N_A p_A + N_B p_B}{N_A + N_B}$$

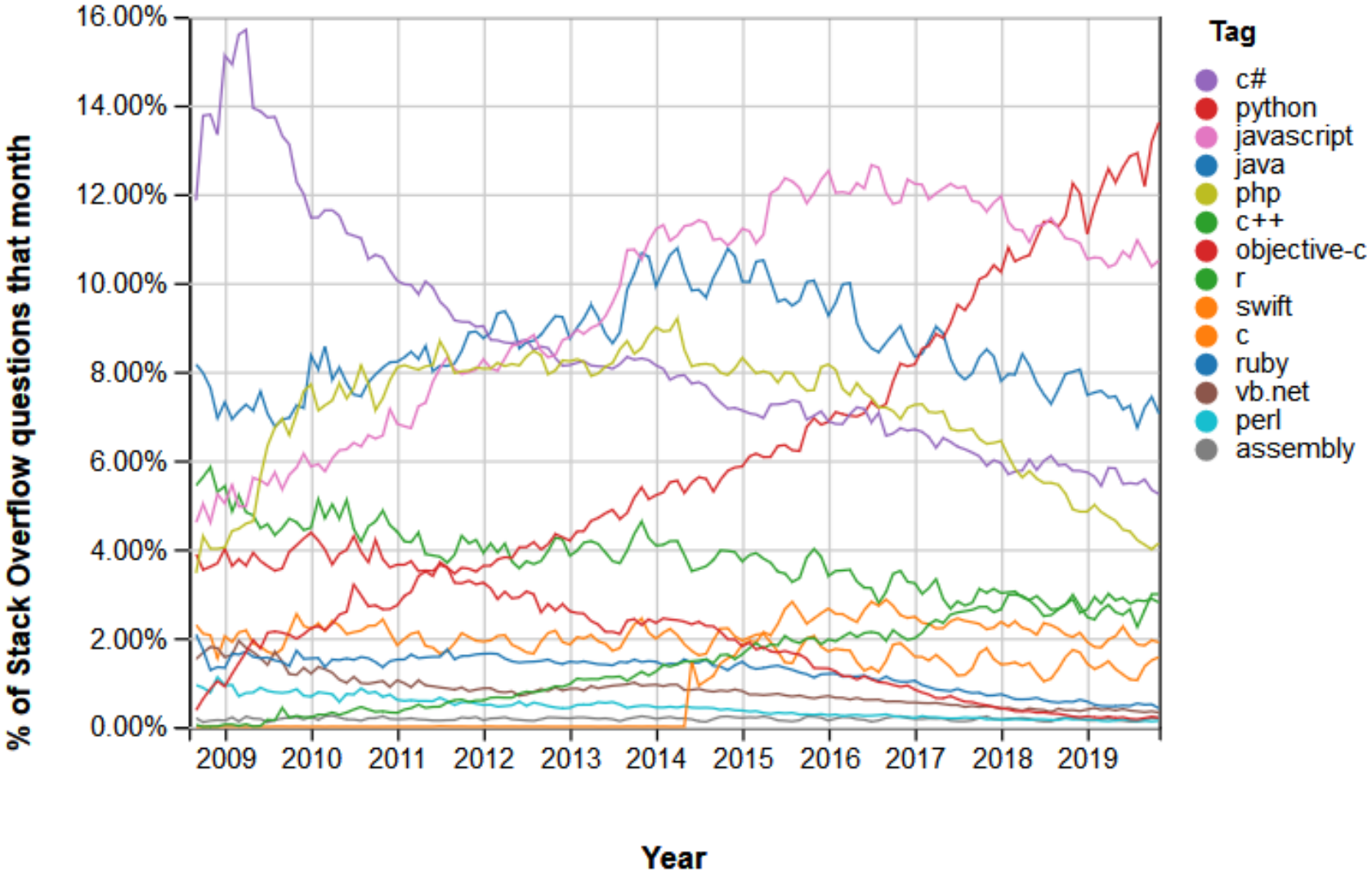
- We assume that null hypothesis is true and two samples are from the same distribution.

$$z = \frac{(p_A - p_B)}{S_{p_A - p_B}}$$

- We can reject the null hypothesis based on the z-statistic.
- I will upload **abtest.csv** for a A/B testing assignment after we talk more about Python and Pandas.

Python

Why Python? Stack Overflow Trend

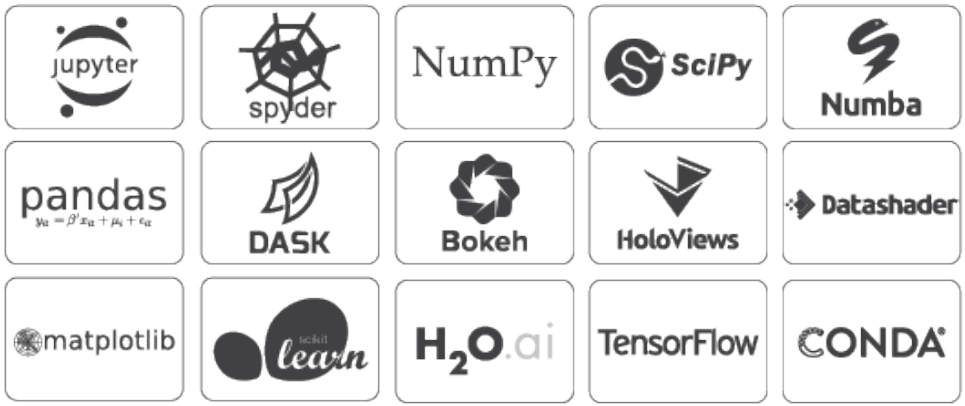


Sources: Stack Overflow Trends

Python

- **Python** is an interpreted, high-level, general-purpose programming language first released 30 years ago.
- **Python** itself was not specifically designed with data analysis or scientific computing in mind.
- The usefulness of Python for data science stems primarily from the large and active ecosystem of **third-party packages**.
- There are different versions of Python. In this class, we will focus on Python 3.7 version.

How to Install Python?



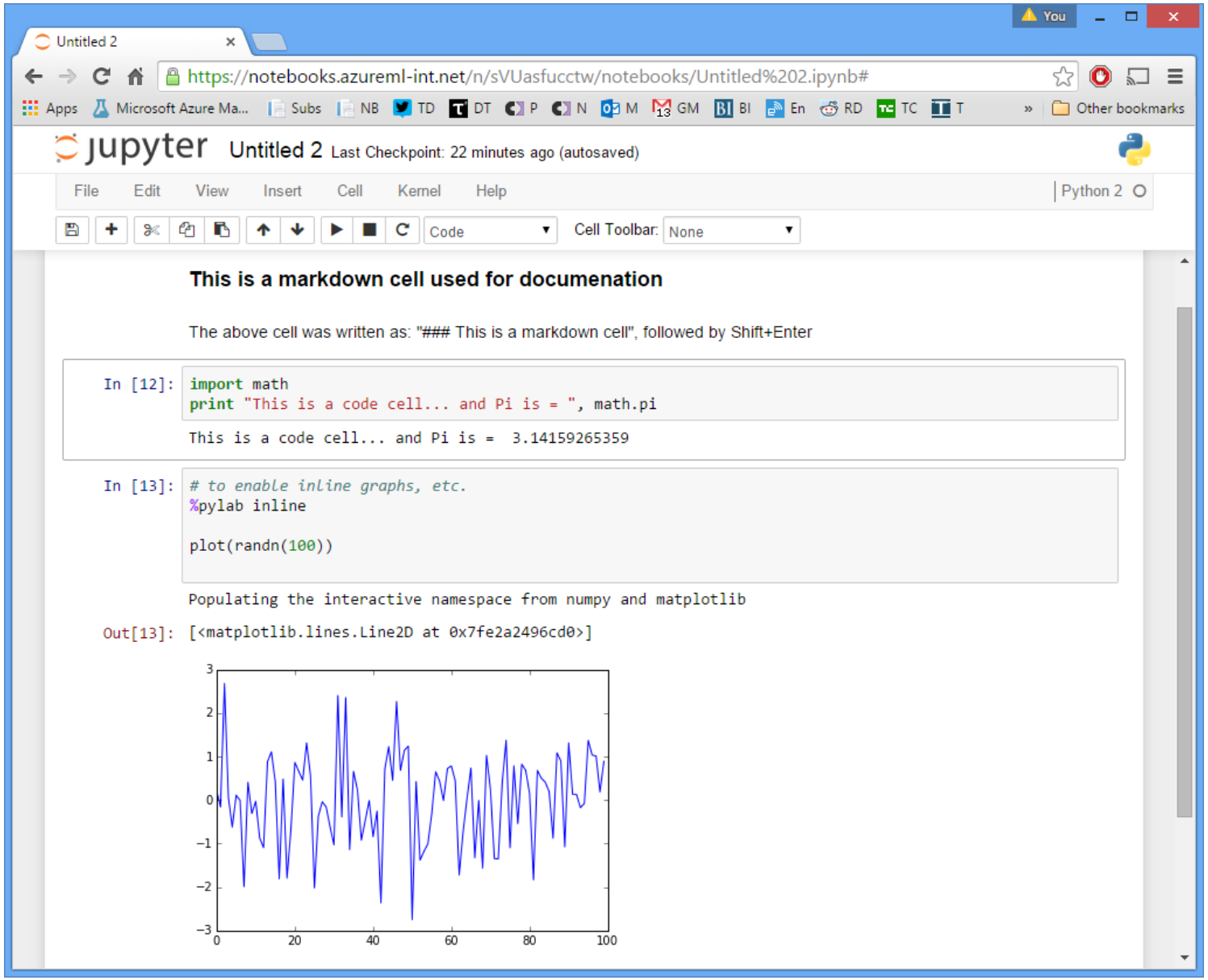
- https://www.anaconda.com/distribution/#download-section
- Important points:
  - Install the latest version. BTW, we use Python 3.8 version.
  - If you are asked during the installation process whether you'd like to make Anaconda your default Python installation, say yes.
  - Otherwise, you can accept all of the defaults.

Integrated Development Environment

Jupyter Notebooks

- **Jupyter notebooks** are one of the many possible ways to interact with Python and the scientific libraries.
- A browser-based interface supporting **Julia**, **Python**, and **R**:
  - The ability to write and execute commands.
  - Formatted output in the browser, including tables, figures, animation, etc.
  - The option to mix in formatted text and mathematical expressions.
- Because of these possibilities, Jupyter is fast turning into a major player in the scientific computing ecosystem.

Jupyter Notebooks Environment



Spyder

- If you are switching from **Matlab** or **Rstudio** to Python; **Spyder** is the way to go, It very intuitive for scientific computing.
- It is integrated into Anaconda. So, you don't need to install it separately.
- I like Spyder because of its **variable exploring** feature.
- Spyder feels more basic than other IDEs. It is not very fancy. But it can definitely get the job done.

Spyder Environment

~/Desktop/Udacity/ML Foundation/Project1 - Bike Share Report - Spyder (Python 3.6)

Editor - /Users/Shiv/Desktop/Udacity/ML Foundation/Project1 - Bike Share Rep...

2nd_panda.py | BayArea-Y3-Summary.csv | example.py

```python
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Thu Jan 11 14:23:25 2018

@author: Shiv
"""
import csv
import statistics as st

subs_data = []
cust_data = []

def dominant_season_for_users(file):

    # Define dictionary.
    with open(file, 'r') as csvfile:

        trip_reader = csv.DictReader(csvfile)

        for row in trip_reader:
            if row['user_type'] == 'Subscriber':
                subs_data.append(row['month'])
            elif row['user_type'] == 'Customer':
                cust_data.append(row['month'])
            else:
                'Not a valid user.'

        subs_mode = st.mode(subs_data)
        cust_mode = st.mode(cust_data)

        return(subs_mode, cust_mode)
```

Variable explorer

Name	Type	Size	Value
chicago	DataFrame	(72131, 5)	Column names: duration, month, hour, day_of_week, user_type
data	str	1	https://vega.github.io/vega-datasets/data/cars.json
df	DataFrame	(4, 3)	Column names: Genre, Rating, Gender
dur_df	dict	3	{'washington_dur':Series, 'chicago_dur':Series, 'nyc_dur':Series}
nyc	DataFrame	(276798, 5)	Column names: duration, month, hour, day_of_week, user_type
wash_mode	int64	1	6
washington	DataFrame	(66326, 5)	Column names: duration, month, hour, day_of_week, user_type

Help | Variable explorer | File explorer

IPython console

Console 1/A

```
In [39]: data
Out[39]: 'https://vega.github.io/vega-datasets/data/cars.json'

In [40]: chart = alt.Chart(data).mark_point().encode(
    ...: ...             x='Horsepower:Q',
    ...: ...             y='Miles_per_Gallon:Q',
    ...: ...             color='Origin:N',
    ...: ...         )

In [41]: chart


In [42]: |
```

IPython console | History log

Permissions: RW End-of-lines: LF Encoding: UTF-8 Line: 7 Column: 1 Memory: 63 %

PyCharm

- PyCharm is an ideal IDE for professional developers working on large projects.
- Looks very fancy!!!
- There are tons of features. They are very helpful for debugging.
- However, PyCharm is a little bit resource-intensive. If have a computer with a small amount of RAM, you may want to use other lighter options.
- There two versions: community version is free.

PyCharm Environment