

파이썬 라이브러리를 활용한 데이터 분석

툴 스터디 1주차

툴 스터디 소개

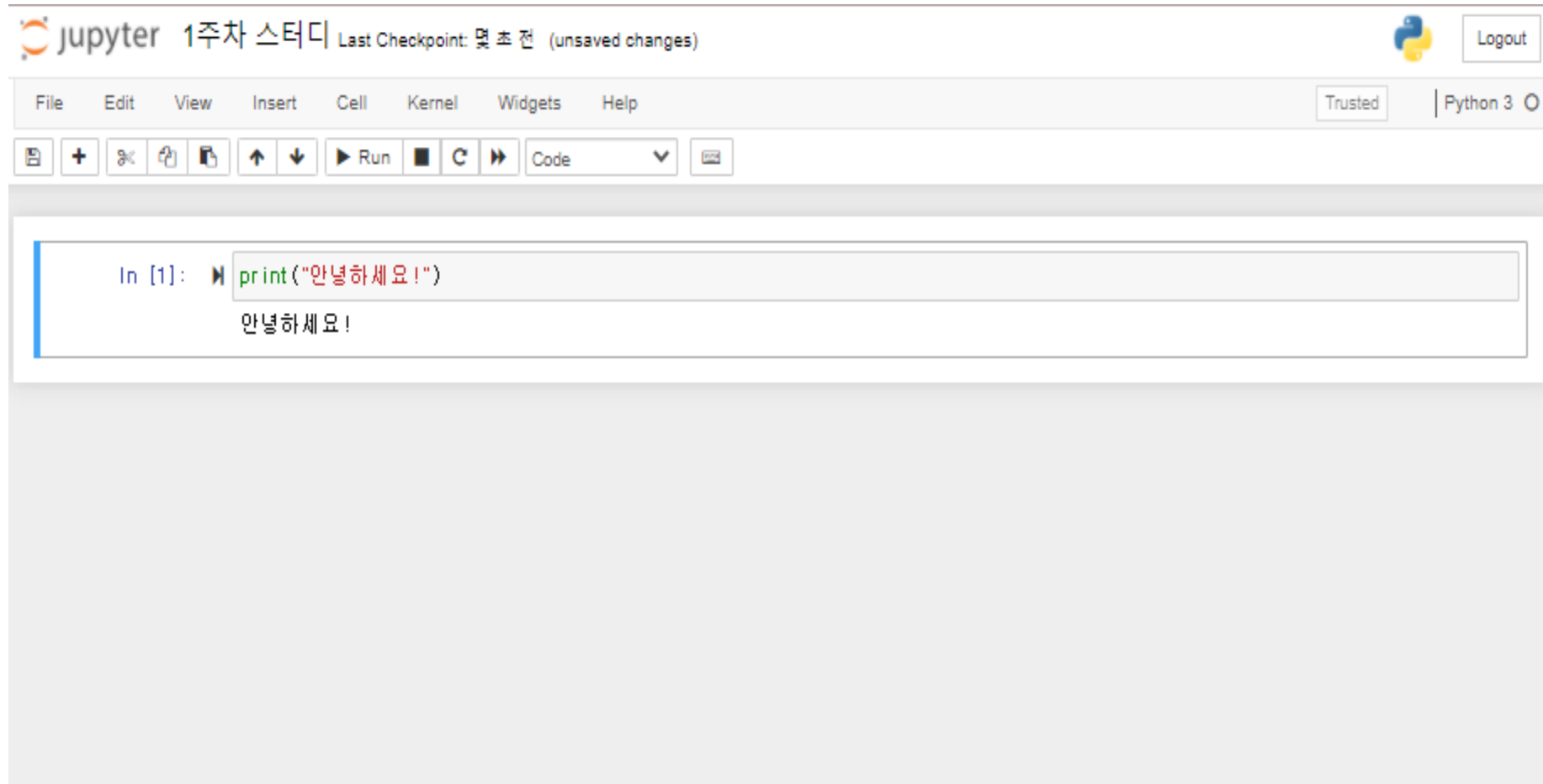
- 캐글의 데이터 분석 예제를 바탕으로, 데이터 분석의 기초가 되는 파이썬의 Pandas, Numpy, Matplotlib 라이브러리를 이해할 수 있습니다.
- 기초 데이터 분석과 파이썬 문법 심화 내용을 병행할 수 있을 것 같습니다.
- 우리 프로젝트의 주요한 데이터 수집 방법은 '웹 크롤링'을 이용한 키워드 분석과, '워드 클라우드'를 이용한 시각화이므로, 이번 스터디를 통해 팀원 분들이 웹 크롤링과 워드 클라우드를 경험할 수 있도록 도울 예정입니다.

툴 스터디 소개

- 커리큘럼 (확정 X)
- 주 1회, 7월-8월 2달 간 총 9회 진행 예정
- 1~2) 타이타닉 예제를 통해 Dataframe을 이해하고,
데이터 전처리와 EDA(탐색적 데이터 분석) 과정 이해하기
- 3) 기초 통계 용어 이해하기
- 4) 웹 크롤링이 무엇인지 알아보고, 프로젝트에서 웹 크롤링을 활용할 방향을 제시하기
- 5) Beautiful Soup를 이용해 웹 데이터를 분석하고, 시각화 하기
- 6) Selenium을 이용해 웹을 다뤄보고, 웹 크롤링을 진행해보기
- 7) Word Cloud를 이용해 수집한 데이터를 시각화하기
- 8) 가설 검정 및 추론, 단순 선형 회귀 분석 이해하기 (확정 x)

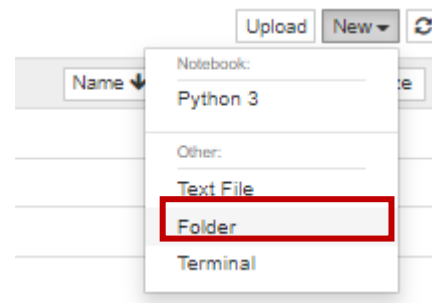
실습 환경 준비

- **쥬피터 노트북 (Jupyter Notebook)**을 준비해주세요!



실습 환경 준비

- 폴더를 만들어주세요.



파이썬에서 '라이브러리'란?

- 라이브러리(Library)는, 특정 기능을 수행하는 함수를 저장해 놓습니다.
- 우리는 매번 코드를 칠 필요 없이, 이 라이브러리를 호출하여 여러 기능을 빠르고, 쉽게, 반복적으로 수행할 수 있습니다.

파이썬의 판다스 (Pandas) 라이브러리란?

- Pandas 라이브러리는 데이터를 분석하고, 조작할 수 있게 해줍니다.
- 우리가 흔히 Excel 파일에서 볼 수 있는, **행(column)**과 **열(row)**로 이뤄진 2차원 구조의 자료를 '데이터 프레임 (Dataframe)'이라고 합니다.

열 (col) 0번째 열, 1번째 열, 2번째 열...

행 (row)
0번째 행
1번째 행
2번째 행
...

	기관명	소계	2013년도 이전	2014년	2015년	2016년
0	강남구	2780	1292	430	584	932
1	강동구	773	379	99	155	377
2	강북구	748	369	120	138	204
3	강서구	884	388	258	184	81
4	관악구	1496	846	260	390	613

- 판다스 라이브러리는 서로 다른 유형의 데이터 (정수형, 문자형, 논리형 True, False 등..)를 '공통의 포맷' 으로 정리해 주기 때문에, 위에서 말한 데이터 프레임 형태의 자료를 분석/조작할 때 자주 사용됩니다.

파이썬의 넘파이 (Numpy) 라이브러리란?

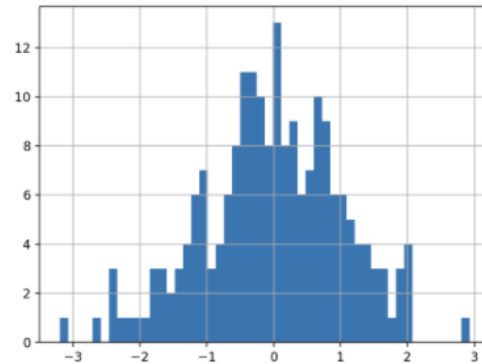
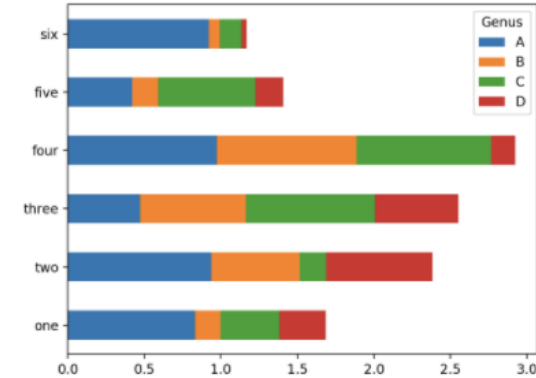
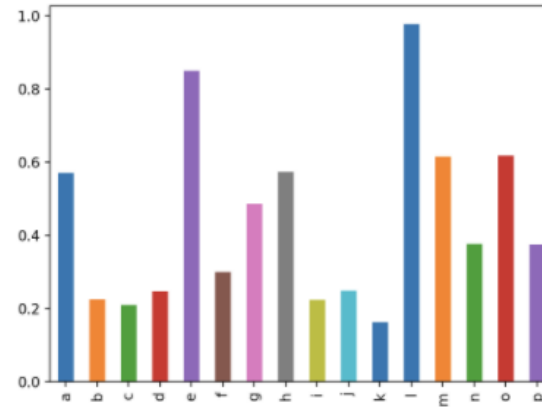
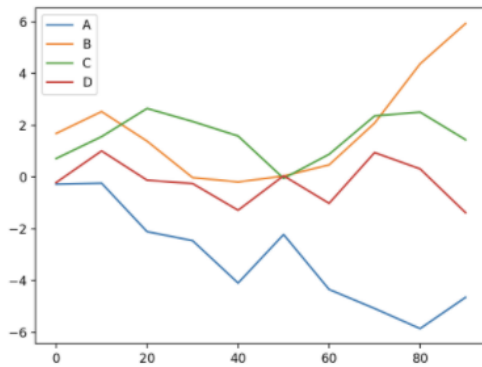
- **Numpy 라이브러리**는 행렬과 같은 다차원 배열을 쉽게 연산하고, 처리할 수 있도록 도와주는 라이브러리입니다.
- 넘파이 라이브러리를 통해서 우리는 복잡한 수식/과정의 수치 계산도 효율적으로 계산할 수 있게 됩니다.

- `15 20 21 22`

- `30 40 23 37`

파이썬의 맷플롯립 (Matplotlib) 라이브러리란?

- Matplotlib 라이브러리는 데이터를 시각화할 수 있는 다양한 기능들이 담긴 라이브러리입니다.



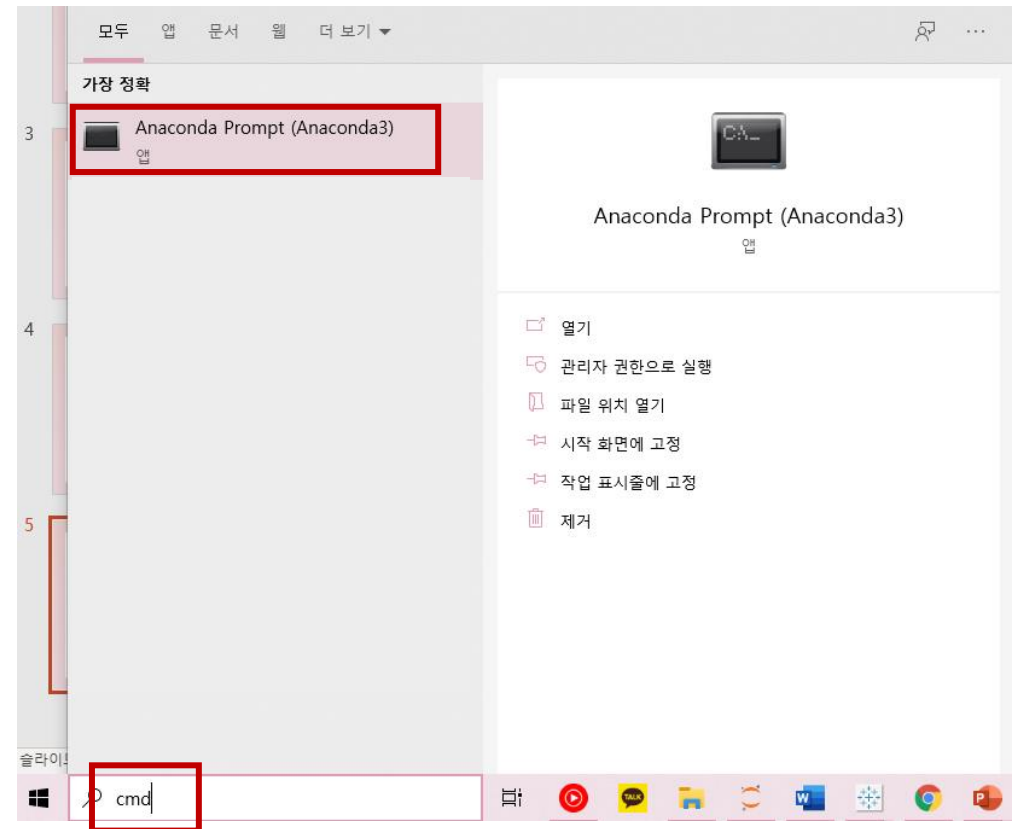
판다스 라이브러리 설치하기

- 1) 윈도우 키를 눌러 시작 메뉴로 이동하기,
- 2) 'cmd'를 검색하여 Anaconda Prompt 실행하기
- 3) 명령 프롬프트 화면에 'pip install pandas' 입력한 뒤,
엔터!

Anaconda Prompt (Anaconda3)

```
(base) C:\Users\holivi>pip install pandas_
```

- 4) 기다리면 설치 완료!



넘파이와 맷플롯립 라이브러리 설치하기

- 명령 프롬프트 화면에 'pip install numpy' 입력한 뒤,

엔터!

```
(base) C:\Users\molivi>pip install numpy
```


- 명령 프롬프트 화면에 'pip install matplotlib' 입력한 뒤,

엔터!

```
(base) C:\Users\molivi>pip install matplotlib
```

판다스 라이브러리 호출하기

- 'as' pd => 앞으로 pandas를 'pd'라는 약어를 통해 호출 가능!

```
In [2]:  import pandas as pd
```

- Pandas.info() pandas.어쩌구
- Pd.info() pd.어쩌구

'타이타닉' 데이터 다운 받기

- 캐글 (Kaggle)은 데이터 분석 경진 대회를 주최하는 플랫폼입니다.
- 캐글에서는 수많은 데이터 분석 예제를 얻을 수 있습니다. <https://www.kaggle.com/>
(캐글에서 데이터를 다운받기 위해서는, Sign In/Register 가 필요합니다!)
- 타이타닉 데이터를 통해서, 우리는 기초 데이터 분석을 배울 수 있게 됩니다.
- <https://www.kaggle.com/c/titanic/data>
- 위 주소에서 타이타닉 데이터를 다운받을 수 있습니다.
- 저희는 조금 더 쉽게 진행하기 위해, 다운받지 않고, 제가 사전에 공유 드린 'titanic.csv' 파일을 바로 이용하도록 하겠습니다.

'타이타닉' 데이터 연결하기

- 아까 만든 폴더에 들어가 'Upload'를 눌러서 'titanic.csv' 파일을 찾아 해당 폴더에 업로드해주기!



- 이렇게 동일한 폴더에 파일을 업로드해주면, 파일을 불러올 때, 해당 파일이 저장된 경로를 따로 찾지 않아도 되어 수월해 집니다!

판다스 라이브러리를 이용한 타이타닉 데이터 분석 Start!

- 그럼 이제 판다스 라이브러리를 이용해 타이타닉 데이터를 살펴봅시다!

다음 시간은...

- 다음 툴 스터디 시간에는 EDA (탐색적 데이터 분석)에 대해 더 자세히 알아보시다!