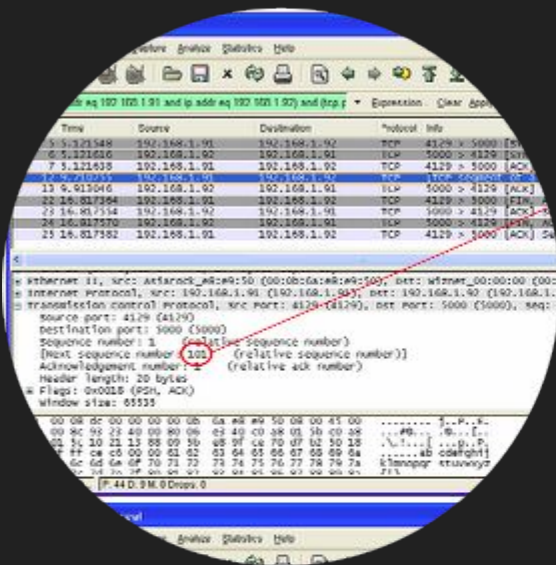


# Python을 활용한 데이터 분석 실습

최규민

# 최규민 소개

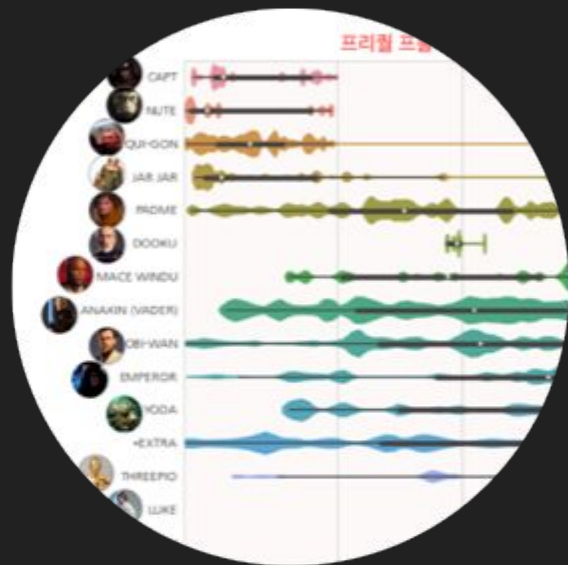
패킷 분석도 해봤고



추천시스템도 만들어보고



데이터 탐색 좋아하여



개발을 즐기는 개발자입니다.

금일 세미나의 자료는 아래 Github에서 다운 받으세요

<https://github.com/goodvc78/t-academy-eda-tutorial>

The screenshot shows the GitHub interface for the repository 'goodvc78/t-academy-eda-tutorial'. At the top, it displays '1 commit', '1 branch', '0 releases', and '1 contributor'. Below this, there are buttons for 'Branch: master', 'New pull request', 'Create new file', 'Upload files', 'Find file', and a green 'Clone or download' button which is highlighted with a red dashed box. A dropdown menu is open for the 'Clone or download' button, showing options to 'Clone with HTTPS' (with a help icon), 'Use SSH', and a text input field containing the URL 'https://github.com/goodvc78/t-academy-eda-tutorial'. Below the URL field are buttons for 'Open in Desktop' and 'Download ZIP'. The repository content is listed below, showing a 'First Commit' with files 'document', 'ipython-nb', and 'README.md'.

File	Commit
document	First Commit
ipython-nb	First Commit
README.md	First Commit



골을 넣기 위해

왼손은  
거들 뿐...

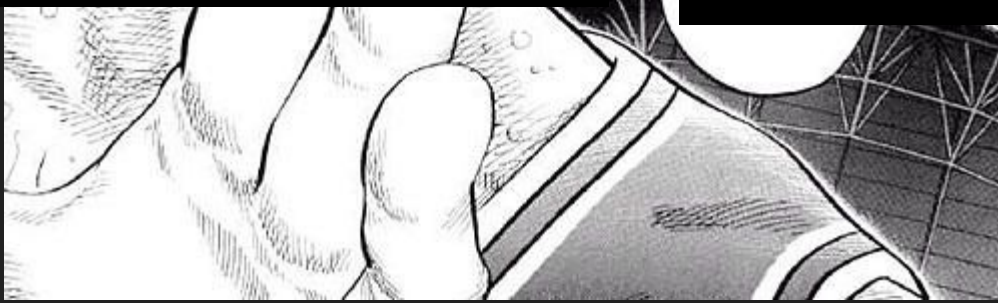






데이터분석을 위해

Python은 거들뿐



# 오늘 할 내용

## 1 탐색적 데이터 분석 사례에 대하여 알아 보기

- 강남 출근길에 정자/판교역에 내릴사람 예측하기

1 Hour

## 2 설문 데이터 탐색 해보기

2.a 데이터 수집(설문 참여)

2.b 데이터 탐색(구글 요약으로 Stop & Think 해보기)

2.c Cluster하여 실습조 나누기 (4~6명 8개조)

1 Hour

## 3 EDA에 많이 사용되는 Python Pattern 익히기

3.a 환경셋팅

3.b pandas 이해하기

3.c EDA Popular Patterns

3.d EDA Popular Pattern 찾아보기

1 Hour

## 4. 데이터 탐색해보기

4.a 설문데이터 탐색

4.b 지하철 자리앉기 데이터 탐색

4.c Kaggle Titanic 데이터 탐색

3 Hour

# 1. 탐색적 데이터 분석에 대하여 알아보기

## 강남 출근길에 정자/판교역 내릴 사람 예측하기

를 하기위한 탐색적 데이터 분석



## 2. 설문 데이터 탐색 해보기



## 2.a 데이터 수집

- 아래의 설문 조사를 해 주세요

<https://goo.gl/forms/j87iphB3Fu7UQgEX2>

### Python을 이용한 데이터 분석 실습 : 설문 조사

"Python을 활용한 데이터 분석" 실습을 위해 설문조사 데이터를 수집합니다.

\* 필수항목

닉네임 \*

내 답변

# 설문 조사 내용

닉네임 \*

내 답변

직업 \*

- ☐ 대학생/대학원생
- ☐ 취업준비생
- ☐ 재직자
- ☐ 기타: \_\_\_\_\_

성별 \*

- ☐ 남
- ☐ 여

나이 \*

내 답변

복장 \*

- ☐ 평상복
- ☐ 캐주얼정장
- ☐ 정장

상의옷색상 \*

내 답변

하의옷색상 \*

내 답변

신발색상 \*

내 답변

신발종류 \*

- ☐ 구두
- ☐ 러닝화
- ☐ 농구화
- ☐ 스니커즈
- ☐ 단화
- ☐ 슬리퍼
- ☐ 샌들
- ☐ 기타: \_\_\_\_\_

사는곳 \*

- ☐ 강북
- ☐ 강남
- ☐ 강서
- ☐ 강동

전공 \*

내 답변

실습환경OS \*

- ☐ 맥 OS
- ☐ 윈도우 OS
- ☐ 리눅스 OS

# 설문 조사 목적

1. 패션이 비슷한 사람끼리 조별
2. 패션과 다른 특징과 상관 관계가 있는지?
3. 재미있을 듯해서(과정을 즐기자)

## 2.b 데이터 탐색하기

설문조사 링크: [설문조사 응답 결과](#)

요약을 통해 각 항목별 분포 탐색해 보기

## 2.c Clustering 해서 조별 나누기

유사도와 군집화에 대하여 알아보자

Clustering Feature를 패션에 관련된 Feature만 사용

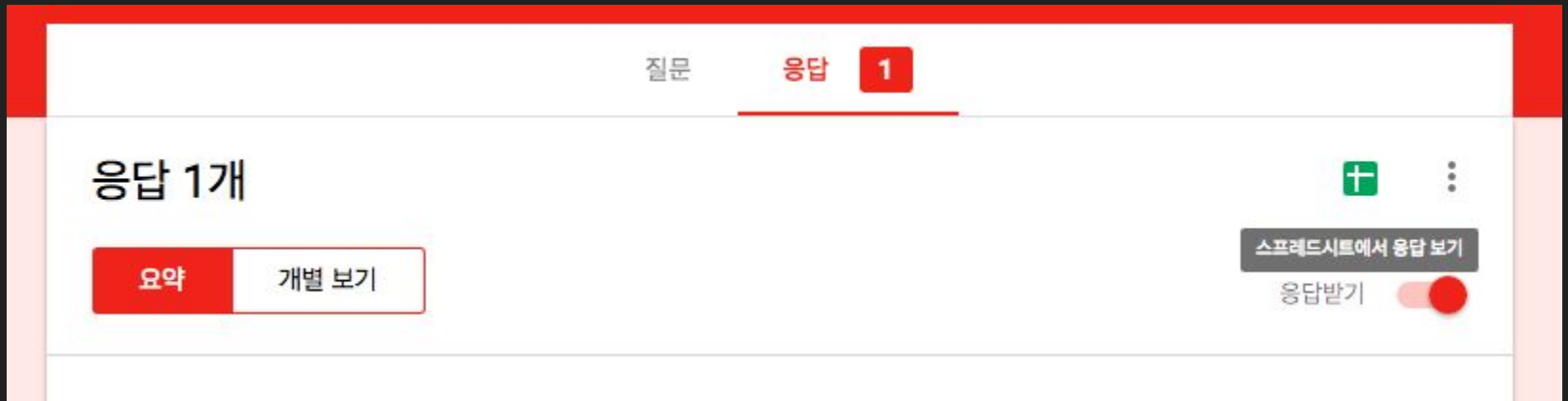
- 상의옷색상
- 하의옷색상
- 하의옷종류
- 신발색상
- 신발종류

계층적 군집화(Hierarchical Clustering) 방식으로 결과를 보고 적당히 Cutting함

결과 탐색해 보고 조별 편성

# 설문 기준으로 유사한 사람끼리 실습조 만들기

설문조사 결과 Export : 응답선택후> 스프레드시트 열기 실행



설문조사 결과의 Value값을 Number로 모두 Encoding

Encoding된 결과로 hierarchical clustering 실행

유사한 사람들 끼리 조별 구성 3~7명 5~8개조

[클러스터링 노트북](#)

# 조별로 Stop & Think 해보기

구글 스프레드 시트로 탐색해 보기

자기 조의 특징에 대하여

특이한 다른 조의 특징에 대하여

Google 문서 시트에 조별로 Stop & Think 추가하기

[https://docs.google.com/presentation/d/1IQoTxbyWD\\_pmg\\_UD92jubSWkr8wlla2ryBFnJD11Mj8/edit?usp=sharing](https://docs.google.com/presentation/d/1IQoTxbyWD_pmg_UD92jubSWkr8wlla2ryBFnJD11Mj8/edit?usp=sharing)



### 3. EDA에서 많이 사용하는 Python Pattern 익히기

## 3.a 환경 셋팅하기

### 설치 패키지

- Python 3.6
- IPython notebook

- 설치할 패키지

Pandas, Numpy

Scikit-learn, Scipy

Matplotlib, Seaborn

- 실습 소스 저장소

github : <https://github.com/goodvc78/t-academy-eda-tutorial>

- Anaconda 설치 추천

사이트 : <https://www.continuum.io/>

설치 참조 : <https://medium.com/@younggun/anaconda-fe67e9c9709d>

# 주로 사용하는 Python 패키지

Pandas : Pannel Data Analysis Package

- 참고: <http://www.slideshare.net/maikroeder/pandas-16424935>

Numpy : Numeric Python Package

Matplotlib / Seaborn : Plotting Package

Scipy : Science Python Package

- dependency : blas, lapack

Scikit-Learn : Machine Learning Python Package

# 자주 사용하는 IPython 단축키

- `ctrl+enter` : 현재 셀 실행
- `shift+enter` : 현재 셀 실행 후 다음 셀 이동
- `alt+enter` : 현재 셀 실행 후 신규 셀 생성
- `Tap` : 자동완성
- `shift + tap` : 현재 커서의 instance의 doc-string 보기
- `shift + ctrl + '-'` : 현재 커서에서 셀 나누기
- `{셀을 여러개 선택후} + 'm'` : 선택된 셀 합치기

## 3.b pandas 이해하기

- Series : 1차원 Array와 비슷한 객체

```
In [27]: pd.Series(np.random.randn(5))
```

Out[27]:

0	-0.564404
1	-0.476555
2	0.240498
3	-0.278255
4	0.747916

dtype: float64

**Index**      **Value**

- DataFrame : 2차원 행렬 객체

```
In [50]: ds = pd.read_csv('./data/devview2015_session_info.csv')
ds.head()
```

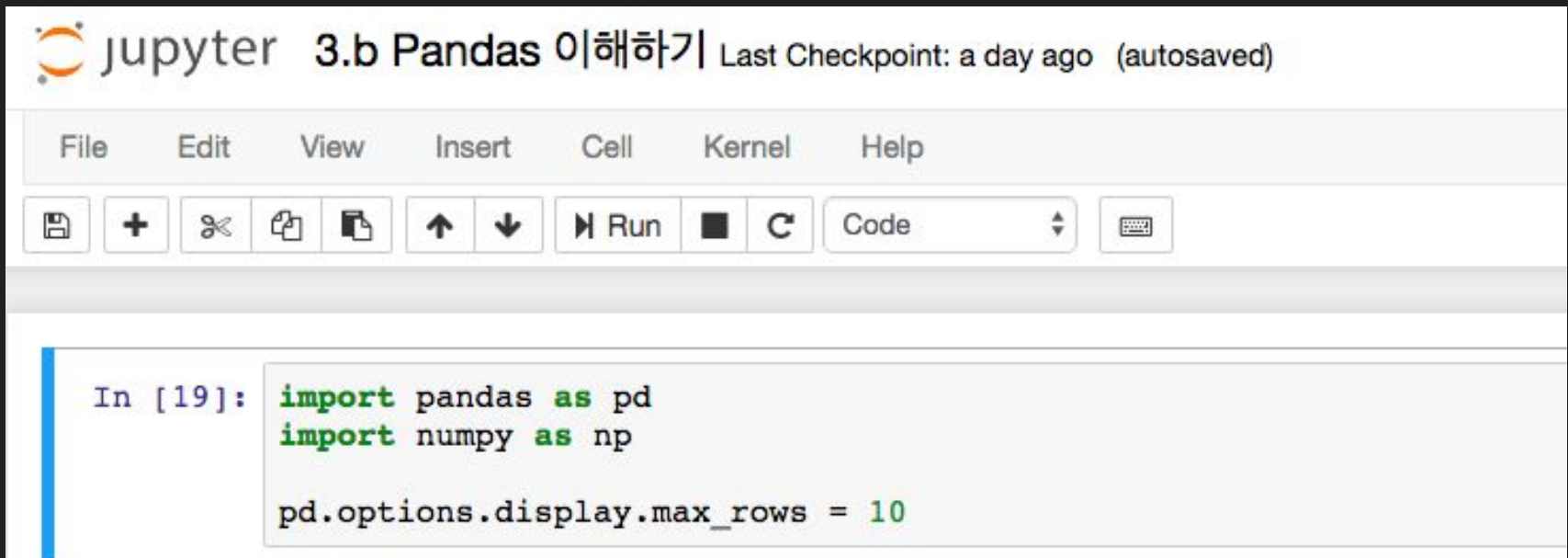
Out[50]:

	day	time	zone	title	session_id
0	Day1	10:00	T1	네이버 효과톤은 어떻게 만들어졌나?	115
1	Day1	10:00	T2	실전 스위프트 프로그래밍	95
2	Day1	10:00	T3	Developing Android Libraries: Lessons from Realim	119
3	Day1	10:00	T4	DRC-HUBO: Technical Review	114
4	Day1	11:00	T1	네이버 효과톤 구현 이야기	89

**Index**      **Column**      **Value**

## 3.b pandas 이해하기

<http://localhost:8888/notebooks/t-academy-eda-tutorial/3.b%20Pandas%20%EC%9D%B4%ED%95%B4%ED%95%98%EA%B8%B0.ipynb>



The image shows a Jupyter Notebook interface. At the top, the title bar reads "jupyter 3.b Pandas 이해하기" followed by "Last Checkpoint: a day ago (autosaved)". Below the title bar is a menu bar with "File", "Edit", "View", "Insert", "Cell", "Kernel", and "Help". Under the menu bar is a toolbar with icons for saving, adding, deleting, copying, pasting, undo, redo, and running code. The main area contains a code cell with the following text:

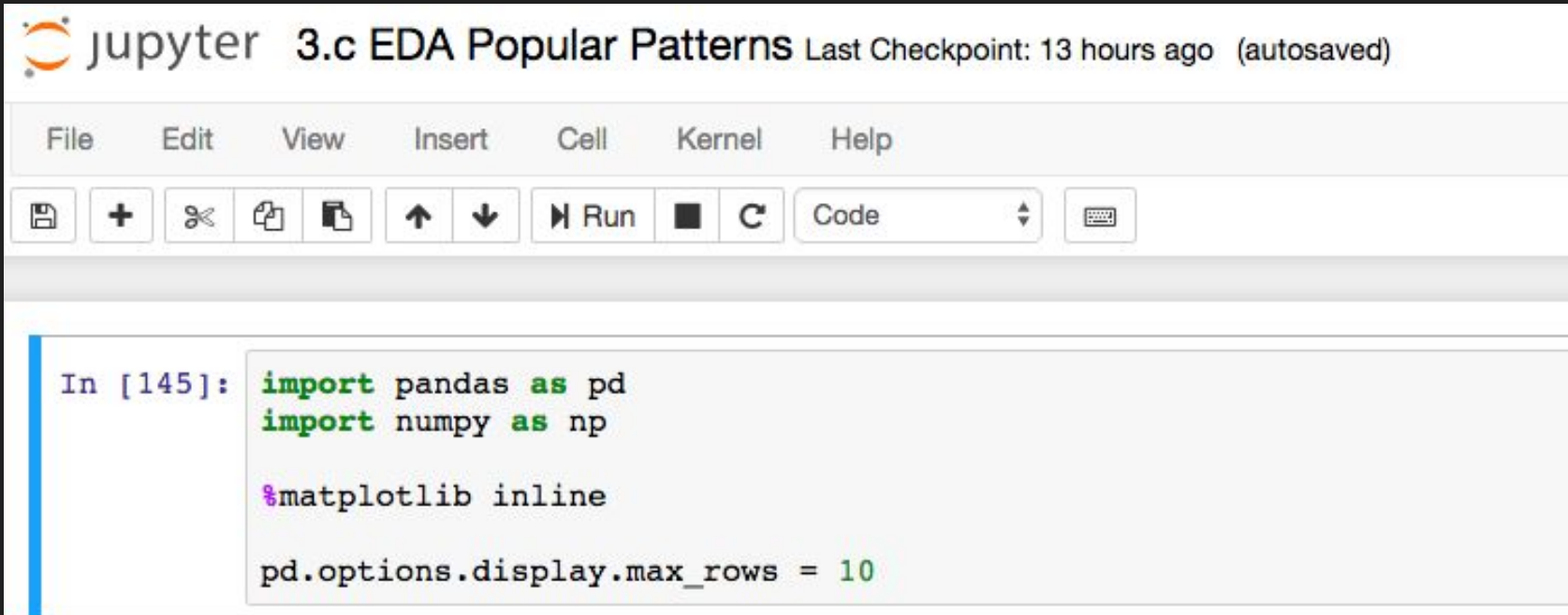
```
In [19]: import pandas as pd
import numpy as np

pd.options.display.max_rows = 10
```

**Series 객체 생성해 보기**

## 3.c EDA Popular Patterns

<http://localhost:8888/notebooks/t-academy-eda-tutorial/3.c%20EDA%20Popular%20Patterns.ipynb>



The image shows a Jupyter Notebook interface. The title bar reads "jupyter 3.c EDA Popular Patterns" and "Last Checkpoint: 13 hours ago (autosaved)". Below the title bar is a menu bar with "File", "Edit", "View", "Insert", "Cell", "Kernel", and "Help". Under the menu bar is a toolbar with icons for saving, adding, deleting, copying, pasting, undo, redo, running, and a dropdown menu currently set to "Code". The main area contains a code cell with the following text:

```
In [145]: import pandas as pd
import numpy as np

%matplotlib inline

pd.options.display.max_rows = 10
```

### 실제 파일의 샘플 확인

- 매직 명령어로 'Head' 명령어 실행하기
- tap을 통한 자동 완성



### 3.d EDA Popular Patterns 찾아보기

Kaggle의 Kernel 많이 쓰이는 Pattern 찾아보기

- <https://www.kaggle.com/kernels>

찾아낸 많이 쓰이는 Pattern을 google-문서에  
기입하기

- [결과2. EDA Popular patterns](#)

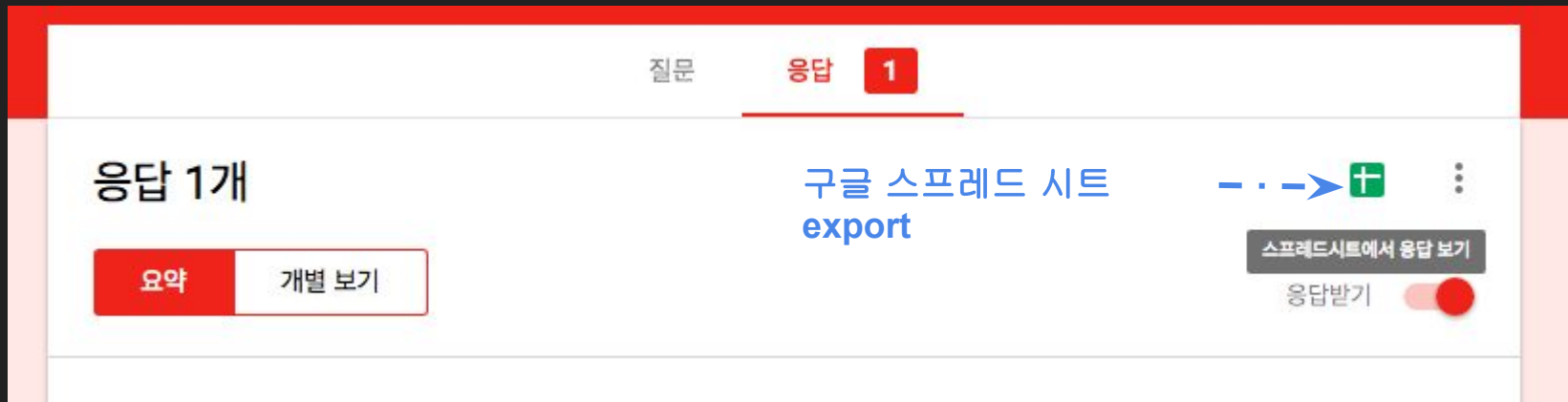
생성된 EDA Popular pattern을 다같이 리뷰

## 4. 데이터 탐색해 보기

## 4.a 설문 조사 데이터 탐색해 보기

- 데이터 수집

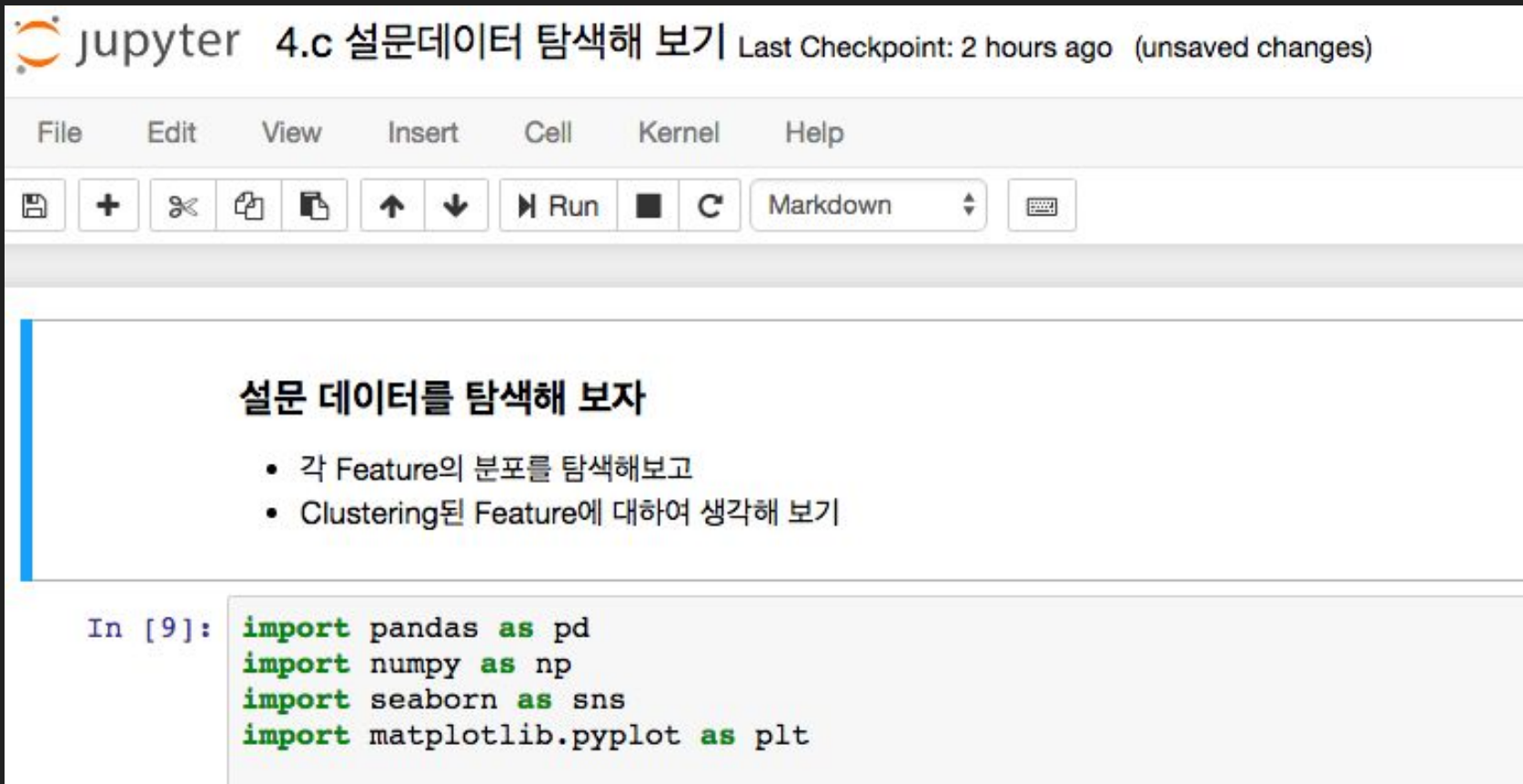
[https://docs.google.com/forms/d/1MG1S3ykUX6\\_4VwQg24pQTrM3nUuRWZdfBBvy-wJWzaw/edit?usp=sharing](https://docs.google.com/forms/d/1MG1S3ykUX6_4VwQg24pQTrM3nUuRWZdfBBvy-wJWzaw/edit?usp=sharing) 접속



- 구글 스프레드 시트 > 파일 > 다른이름으로 저장 > csv 파일 생성
- ./resource/survey.csv

## 4.a 설문 조사 데이터 탐색해 보기

- “4.c 설문데이터 탐색해보기” download
- 간단히 Feature 분포와 clustering을 해서 보자



The screenshot shows a Jupyter Notebook window titled "4.c 설문데이터 탐색해 보기" with a status bar indicating "Last Checkpoint: 2 hours ago (unsaved changes)". The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Help) and a toolbar with icons for saving, adding, deleting, copying, pasting, undo, redo, running, and a dropdown menu currently set to "Markdown".

The first cell is a markdown cell containing the following text:

**설문 데이터를 탐색해 보자**

- 각 Feature의 분포를 탐색해보고
- Clustering된 Feature에 대하여 생각해 보기

The second cell is a code cell with the following Python code:

```
In [9]: import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

## 조별로 데이터 탐색

탐색한 결과 중 가장 의미 있는 결과를  
결과1 아래 문서에 추가

결과1. 설문 데이터 탐색결과

## 4.b 지하철 자리앉기 데이터 탐색해 보기

- 데이터 수집  
./ipython-nb/predict-get-off-station/01-데이터%20수집.ipynb
- 데이터 탐색  
./ipython-nb/predict-get-off-station/01-데이터%20탐색.ipynb
- 데이터 예측 ./ipython-nb/predict-get-off-station/03-지하철%20승객%20하차%20예측하기%20-%20예측하기%20.ipynb

## 4.c Kaggle Titanic 데이터셋 탐색해 보기

- Titanic 데이터셋 EDA 사례 리뷰 : 링크  
<https://www.kaggle.com/c/titanic>
- 조별로 Kernel선택 후 자신의 노트북 또는 Kaggle에서 Fork하여 실행 및 결과 리뷰 (Titanic이 아니어도됨)
- 조별로 탐색한 결과에 대하여 토론하기
  - 5분 발표 / 5분 질답
- Kernel 선택하는 방법
  - Kernel 메뉴에서 인기 커널 선택
  - 자신이 좋아하는 competition 선택 > Kernel 선택



감사합니다.

최규민 (goodvc78@gmail.com)