hci+d lab.

Week 03 • 소셜네트워크 데이터마이닝과 분석

Python Crash Course 2

Joonhwan Lee
human-computer interaction + design lab.

오늘 다룰 내용

- Flow Control
- File I/O
- Data Processing

Data Processing

Data Analysis Process

Data Analysis Process

- Question Phase
  - Characteristics of students who finish MOOC lectures
  - Age and gender distribution of people who spend money in Gangnam area

Data Analysis Process

- Wrangling Phase
  - Data acquisition - where to get data to answer the questions
  - Data cleaning - (in most case) data need to be cleaned - we spend most of our time for this…(80~90%)

Data Analysis Process

- Explore Phase
  - Build intuition by exploratory data analysis
    - information visualization
    - find patterns

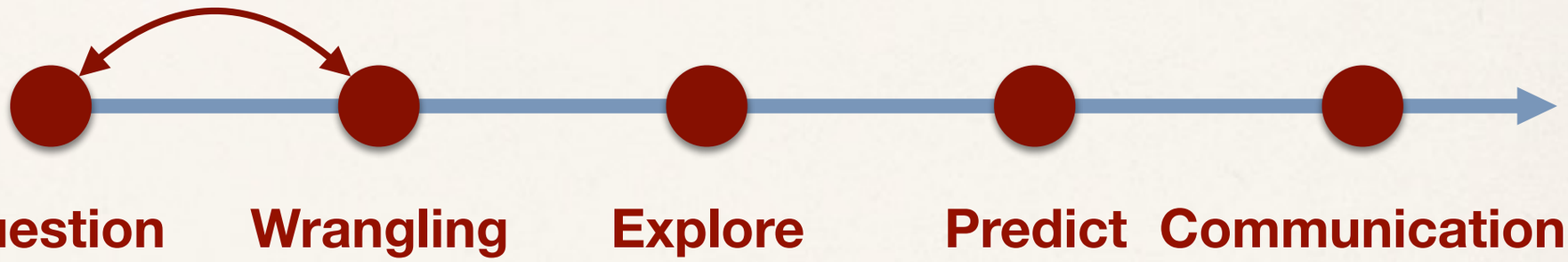
Data Analysis Process

- Prediction Phase
  - Predict results of out question
    - eg. Age and gender distribution of people who spend money in Gangnam area => According to our data analysis, 20-30 women spend more money in this area. => marketing insights
  - Usually requires statistics or machine learning

Data Analysis Process

- Communication Phase
- Data Journalisms
- Blog Posts
- Data Visualizations
- Papers

Data Analysis Process



Question Wrangling Explore Predict Communication

Data Acquisition

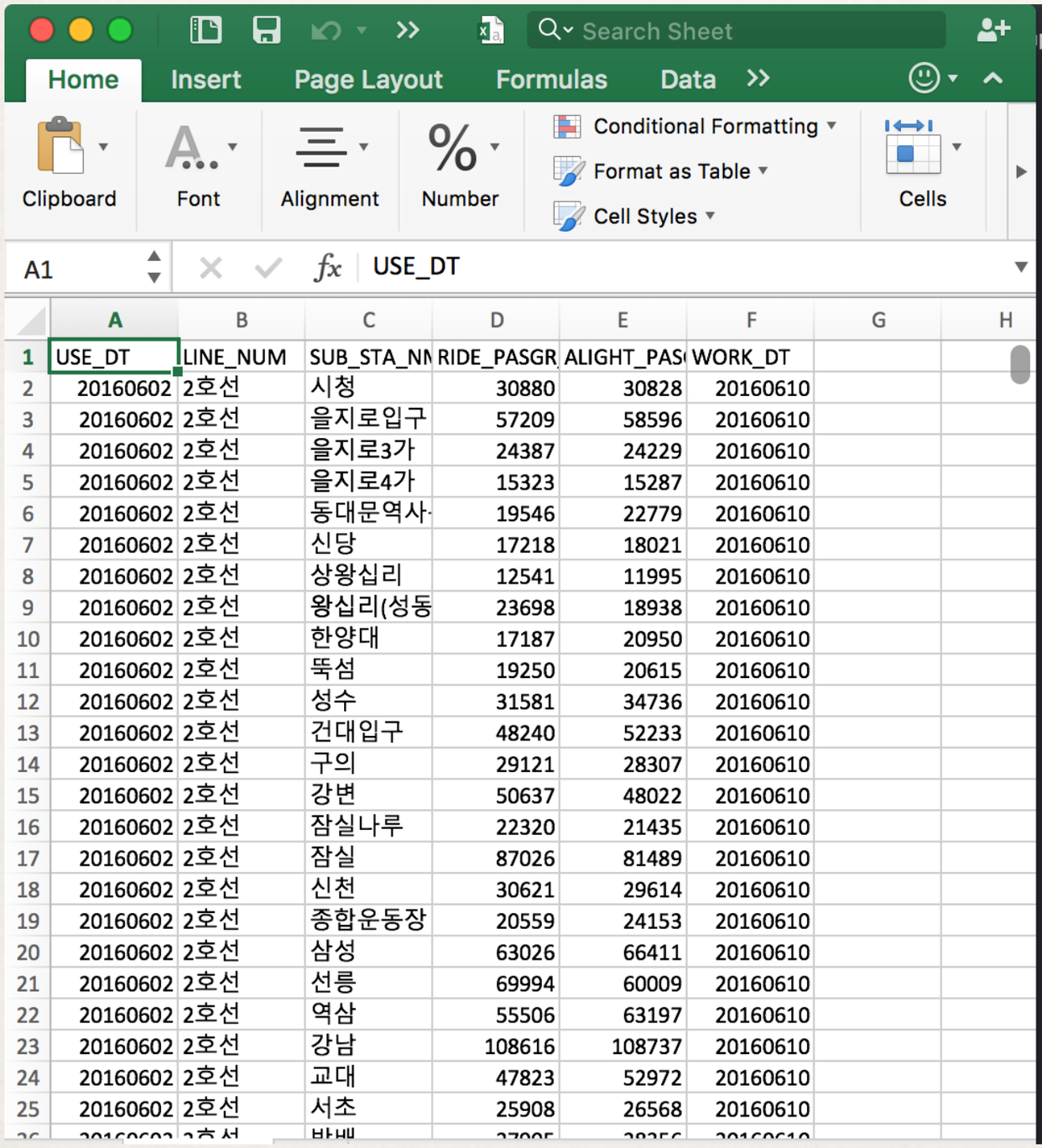
- **Downloading files**
- Accessing an API
- Scraping a web page

→ will do these later

Data Format

- CSV: Comma Separated Values
  - data column separated by comma
  - text file format (xls is binary format) → can read from text editors

Data Format



	A	B	C	D	E	F	G	H
1	USE_DT	LINE_NUM	SUB_STA_NM	RIDE_PASGR	ALIGHT_PAS	WORK_DT		
2	20160602	2호선	시청	30880	30828	20160610		
3	20160602	2호선	을지로입구	57209	58596	20160610		
4	20160602	2호선	을지로3가	24387	24229	20160610		
5	20160602	2호선	을지로4가	15323	15287	20160610		
6	20160602	2호선	동대문역사	19546	22779	20160610		
7	20160602	2호선	신당	17218	18021	20160610		
8	20160602	2호선	상왕십리	12541	11995	20160610		
9	20160602	2호선	왕십리(성동	23698	18938	20160610		
10	20160602	2호선	한양대	17187	20950	20160610		
11	20160602	2호선	뚝섬	19250	20615	20160610		
12	20160602	2호선	성수	31581	34736	20160610		
13	20160602	2호선	건대입구	48240	52233	20160610		
14	20160602	2호선	구의	29121	28307	20160610		
15	20160602	2호선	강변	50637	48022	20160610		
16	20160602	2호선	잠실나루	22320	21435	20160610		
17	20160602	2호선	잠실	87026	81489	20160610		
18	20160602	2호선	신천	30621	29614	20160610		
19	20160602	2호선	종합운동장	20559	24153	20160610		
20	20160602	2호선	삼성	63026	66411	20160610		
21	20160602	2호선	선릉	69994	60009	20160610		
22	20160602	2호선	역삼	55506	63197	20160610		
23	20160602	2호선	강남	108616	108737	20160610		
24	20160602	2호선	교대	47823	52972	20160610		
25	20160602	2호선	서초	25908	26568	20160610		

ple-2.csv

```
USE_DT,LINE_NUM,SUB_STA_NM,RIDE_PASGR_NUM,ALIGHT_PASGR_NUM,
  WORK_DT
20160602,2호선,시청,30880,30828,20160610
20160602,2호선,을지로입구,57209,58596,20160610
20160602,2호선,을지로3가,24387,24229,20160610
20160602,2호선,을지로4가,15323,15287,20160610
20160602,2호선,동대문역사문화공원,19546,22779,20160610
20160602,2호선,신당,17218,18021,20160610
20160602,2호선,상왕십리,12541,11995,20160610
20160602,2호선,왕십리(성동구청),23698,18938,20160610
20160602,2호선,한양대,17187,20950,20160610
20160602,2호선,뚝섬,19250,20615,20160610
20160602,2호선,성수,31581,34736,20160610
20160602,2호선,건대입구,48240,52233,20160610
20160602,2호선,구의,29121,28307,20160610
20160602,2호선,강변,50637,48022,20160610
20160602,2호선,잠실나루,22320,21435,20160610
20160602,2호선,잠실,87026,81489,20160610
20160602,2호선,신천,30621,29614,20160610
20160602,2호선,종합운동장,20559,24153,20160610
20160602,2호선,삼성,63026,66411,20160610
20160602,2호선,선릉,69994,60009,20160610
20160602,2호선,역삼,55506,63197,20160610
20160602,2호선,강남,108616,108737,20160610
20160602,2호선,교대,47823,52972,20160610
20160602,2호선,서초,25908,26568,20160610
```

What are we going to do today?

- CSV import
- Fix Data Type
- Understand Data through Exploration
- Data Filtering
- Add Key(Column) to the Data

Data & Code

- Modified from “Introduction to Data Analysis” course at Udacity.
- Using their login data.
  - Data description included.

Assignment #2-2: Python Crash Course 2

hci+d lab.

Assignment #2-2: Python Crash Course 2

- Exercise 문제
- 다음과 같은 포맷으로 제출
  - 제출 방법: GitHub (다음 주 일요일 3/31 자정까지)
  - 파일 이름: A2-2-학번-이름(영어로).iphynb 형식으로
    (예: A2-2-13403-999-jiyoon.iphynb)

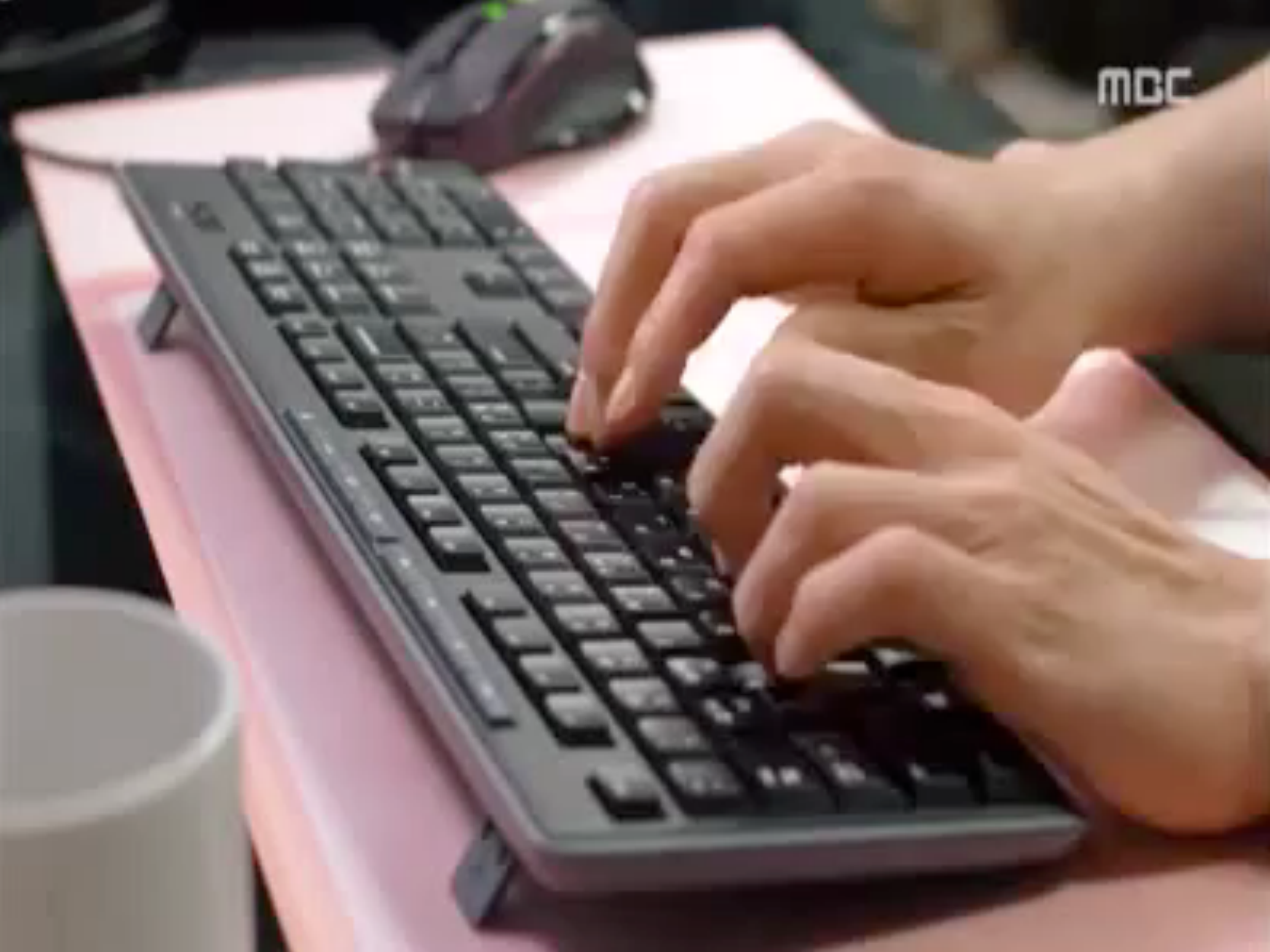
Assignment #2-2: Python Crash Course 2

- #1. 다음을 출력하는 프로그램을 작성하시오.
  - 100 little monkeys jumping on the bed.
    One fell off and broke his(or her) head.
    Mama(or Daddy) called the doctor and the doctor said,
    "No more monkeys jumping on the bed!"
    ...
    ...
    1 little monkey jumping on the bed.
    One fell off and broke his(or her) head.
    Mama(or Daddy) called the doctor and the doctor said,
    "No more monkeys jumping on the bed!"
  - Mama 혹은 Daddy를 랜덤하게 선택하게 하자.

Assignment #2-2: Python Crash Course 2

- #2. 시작하는 연도와 끝나는 연도를 물어보는 프로그램을 작성 하고, 그 사이에 있는 모든 윤년을 출력하라.
  (시작하는 연도와 끝나는 연도가 윤년이면 그 해도 출력)
  - 윤년은 4로 나누어짐 (예: 1984, 2012)
  - 100으로 나누어지면 윤년이 아님 (예: 1800, 1900)
  - 윤년은 400으로 나누어짐 (예: 1600, 2000)

  > Pick a starting year (like 1973 or something): => 1973
  > Now pick an ending year:
  => 1977
  > Check it out... these years are leap years: => 1976
  - (참고) 윤년 리스트: http://mwultong.blogspot.com/2005/12/18002300-leap-year-list.html


MBC

```
3   }
4
5   //게임 로직 부분
6   public class GuguClass {
7       public static void main(String[] args) {
8           System.out.format("%n                                    < gugu>%n");
9           for (int j = 1; j <= 9; j++) {
10              System.out.println();
11              for (int i = 2; i <= 5; i++) {  // 2~ 5
12                  System.out.format("%d X %d = %2d   ", i, j, i * j);
13              }
14          }
15          System.out.println(); // 줄바꿈
16          for (int j = 1; j <= 9; j++) {
17              System.out.println();
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
```

Assignment #2-2: Python Crash Course 2

- #3. 구구단을 출력하는 프로그램을 작성하고 결과를 txt 파일로 저장하시오.

Questions…?