hci+d lab.

Week 04 • 데이터 저널리즘

Data Processing

Joonhwan Lee
human-computer interaction + design lab.

오늘 다룰 내용

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

Data Processing

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

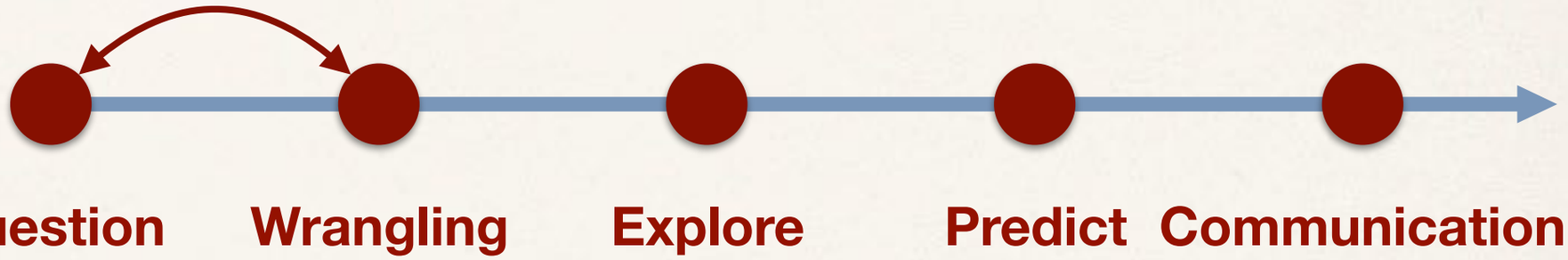
Data Analysis Process

- 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

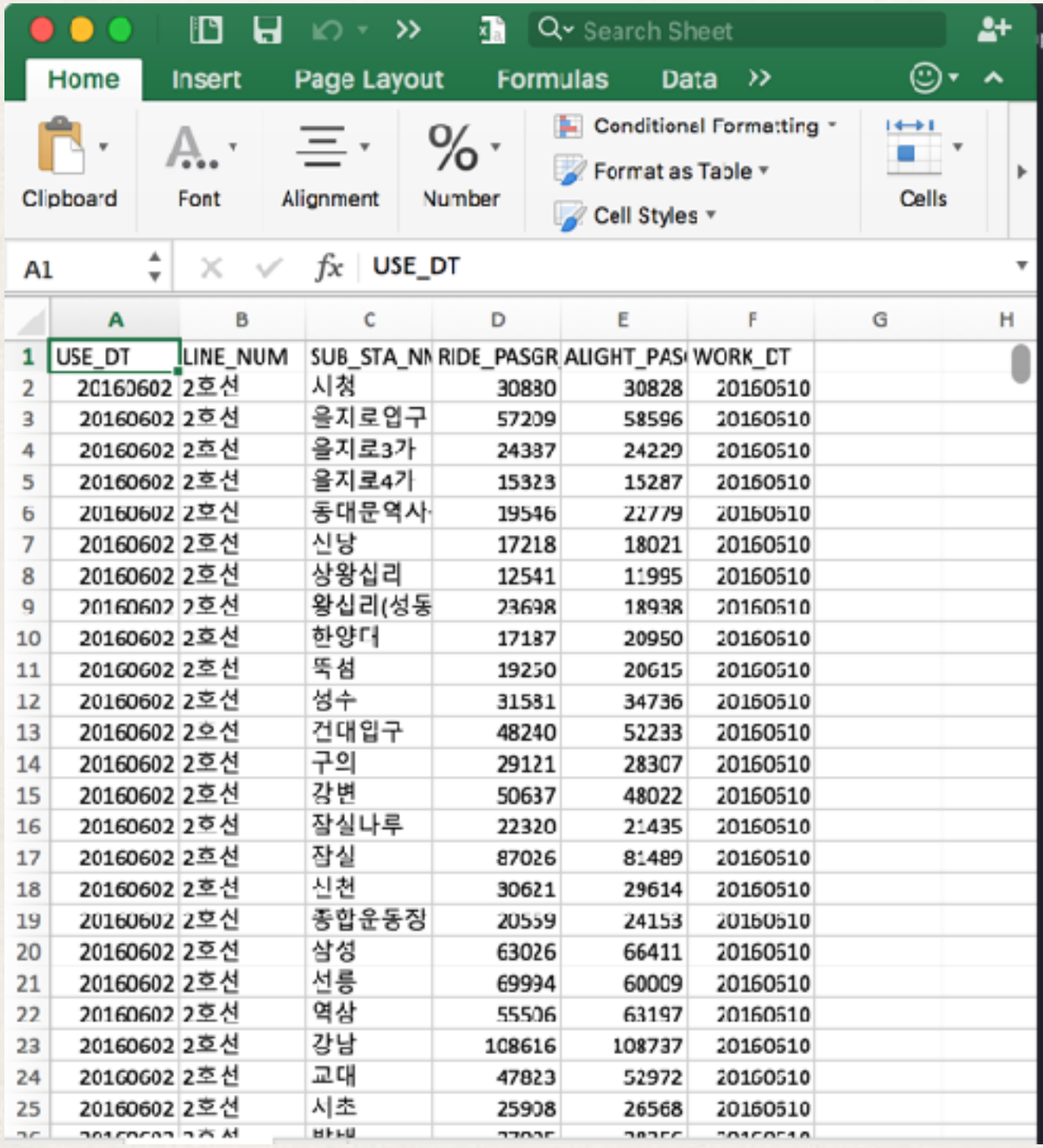
- **Downloading files**
- Accessing an API
- Scraping a web page

→ 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
1	USE_DT	LINE_NUM	SUB_STA_NM	RIDE_PASGR	ALIGHT_PASG	WORK_DT
2	20160602	2호선	시청	30880	30828	20160510
3	20160602	2호선	을지로입구	57209	58596	20160510
4	20160602	2호선	을지로3가	24337	24229	20160510
5	20160602	2호선	을지로4가	15323	15287	20160510
6	20160602	2호선	동대문역사	19546	22779	20160510
7	20160602	2호선	신당	17218	18021	20160510
8	20160602	2호선	상왕십리	12541	11995	20160510
9	20160602	2호선	왕십리(성동	23698	18938	20160510
10	20160602	2호선	한양대	17187	20950	20160510
11	20160602	2호선	뚝섬	19250	20615	20160510
12	20160602	2호선	성수	31581	34736	20160510
13	20160602	2호선	건대입구	48240	52233	20160510
14	20160602	2호선	구의	29121	28307	20160510
15	20160602	2호선	강변	50637	48022	20160510
16	20160602	2호선	잠실나루	22320	21435	20160510
17	20160602	2호선	잠실	87026	81489	20160510
18	20160602	2호선	신천	30621	29614	20160510
19	20160602	2호선	종합운동장	20559	24153	20160510
20	20160602	2호선	삼성	63026	66411	20160510
21	20160602	2호선	선릉	69994	60009	20160510
22	20160602	2호선	역삼	55506	63197	20160510
23	20160602	2호선	강남	108616	108737	20160510
24	20160602	2호선	교대	47823	52972	20160510
25	20160602	2호선	서초	25908	26568	20160510

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 we will 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.

Questions?