데이터 치트 시트 : 크리닝 트릭들

#### **PDF 문서 풀어주기:**

* [Tabula](https://tabula.technology/) : PDF 내 테이블을 CVS파일로 변환해주는 오픈소스 애플리케이션
* [Cometdocs](https://www.cometdocs.com/) : PDF를 CVS 파일로 변환해주는 웹기반 툴(보안 안됨)
* Acrobat: PDF OCR. PDF 문서 내 검색을 할 수 없다면.
* Preview or PDFsplit: PDF 문서 내 페이지를 삭제하거나 분리해준다(보안 안됨)

#### **데이터 조작을 위한 몇몇 유용한 수식들 :**

* =day(cell-reference) 주어진 날짜(cell-reference)의 해당 월의 날짜를 반환해준다
* =month(cell-reference) 주어진 날짜의 월 값을 반환해준다
* =year(cell-reference) 주어진 날짜의 연도 값을 반환해준다
* =text(cell-reference,"dddd") 주어진 날짜의 해당 주의 날을 반환해준다
* 또 유용한 것 : 숫자를 대상으로 한 것처럼, 날짜나 시간을 차감할 수 있다. 예를 들어 A2, A3가 날짜 형태로 존재할 때 A2-A3는 그 사이만큼의 시간 차이를 계산해서 보여준다.

**유용한 키보드 쇼트컷:**

맥에서 수식으로 모두 채우려면, 심지어 빈 셀이 존재하고 있는 경우에라도 :

* 채우고 싶은 수식이 있는 셀을 선택한다.
* shift + command + 아래 방향키를 누른다
  + 이렇게 하면 열의 마지막까지 하이라이트가 될 것이다.
* command + d를 누른다
  + 이렇게 하면 하이라이트 된 셀에 수식이 모두 채워진다.
* 셀을 곧바로 편집하고 싶으면 : fn + f2

#### **코드 북스, 데이터 사전 그리고 메타 데이터 :**

해당 데이터 파일을 이해하는데 필요한 부가 정보. 예를 들어 이 텍스트 파일이 1은 여성을 2는 남성을 의미한다고 설명하거나 또는 해당 데이터가 어떻게 수집됐는지 알려주고 싶은 경우.

#### **데이터 크리닝 트릭들:**

* 순서 바꾸기(Transpose): ‘특수한 붙여넣기' 메뉴에는 한 가지 옵션이 있다. 순서 바꾸기는 행을 열로 만든다든가 혹은 열을 행으로 만든다든가 할 때 당신의 데이터를 한 번에 뒤집어준다.
  + BLS와 같은 정부 데이터 소스를 다룰 때 인기가 있다.
* **데이터 분리(Seperating data)**: Data → split text to columns. 이름과 성, URL처럼 분명한 패턴을 따르는 데이터를 분리할 때 좋다.
* **와일드카드(Wildcards)**: 오타 교정이나 대량 데이터 작업을 위해 수식이나 필터 내에서 사용될 수 있다
  + \* 알 수 없는 문자의 수를 나타낸다
  + ? 한 자짜리 문자를 나타낸준다
  + ~ 다른 와일드카드 문자 중 하나를 추적해 나타낸다
  + 더 많은 옵션이 필요하다면, regex를 사용해보라
* **=EXACT(value1,value2)** 만약 두 텍스트 값이 동일한 상태가 참인지, 그렇지 않은 상태가 거짓인지 확인할 때 사용한다. 복사-붙이기를 할 때 스스로 체크해보는데 유용하다.