

## [과제 5] 시각화와 문서화를 통한 탐색적 데이터 분석

최종 결과물인 RMD파일과 이를 변환한 HTML 파일을 모두 제출해야 함

## 1. 과제의 목적

데이터를 숫자와 문자 그대로만 보는 것이 아니라 이러한 숫자와 문자가 우리에게 주는 메시지를 데이터 분석 및 시각화를 통해 찾는 연습을 하는 것

(데이터 시각화를 통한 메시지 전달의 좋은 예시)

**TED** Ideas worth spreading

Hans Rosling | TED2006  
한스 로슬링이 이제껏 보지 못했던 최고의 통계를 보여준다.

나머지 모든 국가들이 저 구석으로 올라갑니다.

15:04

**Details** About the talk  
**Transcript** 49 languages

이 같은 데이터를 본 적이 없을 것이다. 드라마틱한 이야기 전개와 스포츠 캐스터 같은 열의있는 발표를 통해, 통계 전문가 한스 로슬링이 소위 말하는 개발 도상국에 관한 통계를 완전 해부한다.

This talk was presented at an official TED conference, and was featured by our editors on the home page.

**14,841,229** views

**TED2006** | February 2006

**Related tags**  
Africa  
Asia  
Google  
...

**ABOUT THE SPEAKER**

**Hans Rosling** · Global health expert; data visionary

In Hans Rosling's hands, data sings. Global trends in health and economics come to vivid life. And the big picture of global development -- with some surprisingly good news -- snaps into sharp focus.

[https://www.ted.com/talks/hans\\_rosling\\_the\\_best\\_stats\\_you\\_ve\\_ever\\_seen?language=ko#t-270426](https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen?language=ko#t-270426)

## 2. 주제 선정 및 적합한 데이터셋 다운로드

본인이 본 과제를 수행하면서 하고자 하는 주제를 먼저 설정하고 이를 수행할 수 있는 적절한 데이터셋을 찾아서 다운로드 하시오.

(주제 예시 1) 국가별 COVID-19 발생 동향 분석

(데이터셋이 포함된 웹페이지)

<https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide>

(주제 예시 2) 우리나라의 산업 성장에 대한 다양한 지표들의 연계 분석

(가용 데이터 출처: 통계청)

<http://kostat.go.kr/portal/korea/index.action>

주제에 대한 선정이 어려운 학생들은 Kaggle 웹사이트(<https://www.kaggle.com/>)에서 제시된 데이터셋의 목적을 분석 주제로 선정해도 무방합니다.

The screenshot shows the Kaggle Datasets page. On the left is a sidebar with navigation links: Home, Compete, Data (selected), Notebooks, Communities, Courses, and More. The main content area is titled 'Datasets' and includes a search bar, a 'New Dataset' button, and a section for 'Create Public Datasets'. Below this is a search bar for '63,386 datasets' and a list of public datasets sorted by 'Hottest'. The datasets listed are:

- California Traffic Collision Data from SWITRS** by Alex Gude, updated 19 days ago. Usability 9.7, 1 File (SQLITE), 1 GB, 1 Task. 263 votes, Silver medal.
- Women Entrepreneurship and Labor Force** by babyoda, updated 20 days ago. Usability 10.0, 1 File (CSV), 1 KB, 1 Task. 203 votes, Gold medal.
- Internet news data with readers engagement** by Szymon Janowski, updated 20 days ago. Usability 9.4, 1 File (CSV), 3 MB. 91 votes, Silver medal.
- Credit Card customers** by Sakshi Goyal, updated 22 days ago. Usability 10.0, 1 File (CSV), 379 KB, 2 Tasks. 201 votes, Silver medal.

On the right side, there is a section for 'Open Tasks' with several tasks listed, including 'Trying foods as you grow older', 'Australia News and Economic C...', 'Occupancy Prediction', 'Insights from online pet food cu...', and 'Process condition prediction'.

### 3. 데이터에 대한 탐색적 분석 및 시각화

R Notebook과 ggplot2 package에서 제공하는 기능을 사용하여 최소 10개 이상의 다른 형태의 그래프를 그려보고 이를 통해 알 수 있는 사실을 본인만의 언어로 서술하시오.

ggplot2 package로 도시할 수 있는 그래프의 종류는 아래 웹페이지를 포함해서 직접 다양한 사례를 조사하고 이를 적극적으로 활용하세요.

<http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html>

<https://www.r-graph-gallery.com/ggplot2-package.html>

#### AN OVERVIEW OF **GGPLOT2** POSSIBILITIES

Each section of the gallery provides several examples implemented with **ggplot2**. Here is an overview of my favorite examples:

