**Information Seeking: Possible Uses of Three Datasets**

1. **CIA World Factbook Internet Penetration Table**

CIA World Factbook. (2014). *Country Comparison: Internet Penetration.* Retrieved from <https://www.cia.gov/library/publications/the-world-factbook/rankorder/2153rank.html>

**Terms of Use:**

The factbook’s website states that the data is available for public access.

**Why the dataset is interesting**:

The dataset shows the proportion of the population of countries around the world (217 countries) that use the internet. It can be used in conjunction with information on social media use to assess the social impact of smart devices. The table shows the number of users of the internet but not the percentage of total population, as well as the year that information was collected.

**Potential users and decision makers of the data:**

Any researchers interested in the impact of internet penetration around the world would benefit from data on how many people in each country use the internet.

**Questions about the data:**

* What is the country with most users of the internet?
* What is the country with least users?
* (In conjunction with population data) What is the country with highest/lowest internet penetration?

1. **City of Chicago Department of Streets & Sanitation Graffiti Removal Data**

**Citation**:

City of Chicago. (2012). *Performance Metrics – Streets & Sanitation – Graffiti Removal*. Retrieved from [https://data.cityofchicago.org/Administration-Finance/Performance-Metrics- Streets-Sanitation-Graffiti-Re/qcfn-tiw7](https://data.cityofchicago.org/Administration-Finance/Performance-Metrics-Streets-Sanitation-Graffiti-Re/qcfn-tiw7)

**Terms of Use:**

The dataset is published on the City of Chicago’s data website and is classified as “open data.”

**Why the dataset is interesting:**

The dataset is comprised of 67 rows and 5 columns, and details information on the Department of Street and Sanitation graffiti blasters crew’s vandalism removal service. The data was collected over a period from April 2011 to July 2012 and “tracks the average number of days DSS takes to complete graffiti removal requests per week. Median days to complete requests as well as total number of requests fulfilled per week are available by mousing over columns.” (City of Chicago data website). Such detailed data on how the City tackles vandalism claims is interesting because it gives an idea of the prevalence of graffiti in the city (in some weeks over two thousand requests to remove graffiti are completed) and the number of days taken to complete the requests is a concrete way to track the graffiti blasters crew’s efficiency.

**Potential data users and decision makers of the data:**

Mischievous graffiti artists could use the data to understand how the graffiti blasters crew operates to subvert their vandalism removal efforts. Supervisors in the graffiti blasters crew can use the data to assess the crew’s effectiveness and determine whether more training or different approaches are needed to tackle Chicago’s graffiti problem.

**Questions about the data:**

* How long, on average, does it take the crew to complete a request?
* Is there a time of the month during which there are more requests? A time of the year?
* Does the graffiti blasters crew require more members to complete its task effectively?

1. **Metadata from The Million Song Dataset**

**Citation**:

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. (2011). *The Million Song Dataset*. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011). Retrieved from <http://labrosa.ee.columbia.edu/millionsong/pages/getting-dataset#subset>

**Terms of Use:**

The dataset website’s home page (linked below) states that the data was created under a grant from the National Science Foundation and is freely available to whoever wants to use it.

**Why the dataset is interesting:**

This is a dataset containing the metadata of The Million Song Dataset project, which alone is 300 MB. The original dataset contains data on a million pop songs and is “a freely-available collection of audio features and metadata for a million contemporary popular music tracks.” (Million Song Dataset website home page). The aim of the dataset is to provide a large sample of data on songs on which algorithms (perhaps for song ordering purposes for music streaming) can be tested. Such a large dataset (a million records) requires extensive metadata for it to be easy to understand and manipulate by researchers.

**Potential data users and decision makers of the data:**

Such a large dataset of contemporary popular music is of interest to all stakeholders in the mainstream music industry and particularly developers of applications surrounding music such as music streaming services.

**Questions about the data:**

* How long are the most popular pop songs?
* What is the gender distribution among singers of the top ten most popular songs for a given year?
* Do song titles reflect song content (genre)?