Challenge 2021

Rocky Mountain Poison and Drug Safety center is interested in discovering and identifying patterns of drug use, with particular attention paid to identifying misuse. For example, patterns might describe demographic profiles within a given category of drug, or combinations of drugs frequently used across groups of people, or combinations of drugs that frequently appear together. One goal of these data is to predict future drug misuse cases. These profiles might be used to predict future drug misuse cases, or to develop a short questionnaire (10 or fewer questions) that might help physicians predict future misuse.

Data collection method: the data, as presented, represent a weighted random sample from each of several countries. Although the weights are provided, we recommend that you ignore these weights and regard each sample as a convenience sample. (See notes below on how this assumption affects the scope of inference.) Only use the weights if you are familiar with weighted survey samples. You will NOT be penalized for ignoring weights and will not be compared unfavorably with a team that does. You might be penalized for overstating your claims, and so reading the Scope of inference section below is important.

Proprietary Statement

Data from the RADARS® System Survey of Non-Medical Use of Prescription Drugs Program **are not public.** These data are **proprietary** and are only to be used for the purposes of the American Statistical Association's DataFest. By using this data, you agree to:

1. During participation in DataFest, **store** and **manage** the data **securely** and **privately**.
2. **Erase all data** after your DataFest participation is complete.
3. **Not identify or attempt to identify** the information contained in the dataset, **nor contact** any of the individuals whose information is contained in the dataset.
4. **Comply with** all applicable U.S. federal and state laws and **regulations** relating to the maintenance of the dataset, the safeguarding of the confidentiality of the dataset, and the use and disclosure of the dataset.
5. **Not publish** results of your analysis of the data except that the final products of the competition (video, slide deck, one-page summary) may be displayed on team members' websites and on campus DataFest websites.
6. **Not share** the data with anyone who is not a participant of DataFest.

Table of contents

Data structure

Each file contains a sample from a different country, although there are two files for the U.S.:

country	filename	collection dates	size (rows X columns)
Canada	ca.csv	September 2017 - December 2017	10007 x 185
Germany	de.csv	December 2017 - January 2018	15051 x 156
UK	uk.csv	March 2018 - April 2019	10006 x 187
US18	us18.csv	March 2018 - May 2018	30007 x 199
US19	us19.csv	March 2019-May 2019	29873 x 523

Notes

Each file has its own data dictionary, and the survey questions asked vary from file to file. (Slightly different questions are used in the two US samples, as well. The US19 survey was re-worded in order to be implemented on mobile devices.) The data dictionaries define the variables and the values within each variable and connect them to specific survey questions.

It is sufficient to work within a single file. There is no requirement to merge files or do comparisons across files. In fact, such comparisons should be undertaken with supreme caution since question wording varied.

Please read the Scope of Inference section below to understand the sort of conclusions that can and cannot be made with these data.

Some countries include supporting information files. Each country includes a data dictionary, but the US, for example, the UK includes a file that defines postal codes.

In the US, CA, and UK surveys the order of questions was randomized; in DE they were not. Information about the order in which a respondent was asked the questions is provided within the data files if necessary.

Scope of inference

The sample consists of participants who responded to an on-line request and were paid to participate. A representative sample can be obtained by utilizing the provided probability weights (provided in the variable *WT)* but we strongly recommend that these weights be ignored and, instead, you apply the following limits to your conclusions:

- Data should be treated as a convenience sample.
- Comparisons between countries should be avoided.
- Patterns within countries should be taken as interesting "signals" that point towards areas requiring further investigation, but should not be presented as definitive or generalizable to a larger population beyond the sample. For example, if you find that opioid use is higher in Montana than in Louisiana without using the weights, this would not necessarily mean that the actual prevalence of opioid use is higher in one state than the other. It could be framed as something interesting or speculative that calls for further investigation, though.
- Beware of comparisons across zip-codes. The data are provided at a fairly granular level, but are quite sparse at this level. For example, there may be only one respondent within a given zip code. Spatial visualizations should be interpreted with this in mind.

- A "quick fix" that will make the sample generalizable to the population (which is defined as the country from which the sample was obtained) is to reproduce each row the number of times given by the WT variable. For example, if WT=10, then reproduce that row 10 times. However, this will result in a much larger file that may crash your computer.
- For the U.S. data files, replicating rows only within states allows for valid point estimates of those states, but not valid variance estimates. Therefore, confidence intervals, hypothesis tests, etc. based on the sample with replicated rows will not be valid. Similarly for other large regions.
- If you have worked with weighted samples and are familiar with techniques and software that will correctly account for weights, then you may use these, but you must inform the judges of your methods within both the video and, in more detail, on the one-page summary.

Final products

You will be asked to submit two items for judging. The primary item is a short (6 minute) video detailing your primary findings. The second is a one-page summary of your findings. Please follow these guidelines:

1. **Video:** Time must be less than 6 minutes. We recommend simplicity and suggest you create powerpoint slides (or equivalent) and record a video playing the slides with audio. The video should begin with the team name and the members' names. We encourage visualizations and tables over lists of facts and numbers (which can be delivered via voice).

2. **Summary Paper:** the summary paper should be no more than 1 page, 12 point font, 1 inch margins, single spaced. It should include: team name and members' names, primary questions that were investigated or the general goal of the project, methods used, and quick description of findings. You do not need to include visualizations, but can refer the reader to the video for any tables or graphs. You will not be penalized for your choice, but will be penalized for failing to disclose this information or for reaching a conclusion that is not supported by your choice. If you don't know what to include, then think of this as the script for your presentation.

Advice for getting started

Because of the large number of files, variables, and possible pathways on which a data exploration might embark. We offer some basic advice to help:

1. Begin by reading the data dictionary of the country you're most interested in.
2. Next, discuss with your team a few questions that you find interesting. If you think it's interesting, others probably will, too. It is worthwhile to think ahead about how your questions and your analysis will be perceived during the judging.
3. If your questions *are* interesting, they will probably be too difficult to do in one weekend. So break the problem down into smaller, what might seem like less-interesting, parts. If you can tackle a few of these smaller parts, you can report on the pieces as part of this larger investigation, and report on the progress you made towards the larger question. You don't have to stick with your interesting questions if they don't prove fruitful or if new pathways open up. Still, asking good questions is the place to begin, even though their only use may be to take you to new questions.

4. Think small. A large-scale but vague analysis is less feasible (and, ironically, might be less interesting) than a small-scale, detailed analysis.
5. Avoid complex models until you understand the data well. For example, a common error is to put all of the variables into a single model and then turn the crank and see what comes out. This is never productive.
6. A similar error is to create many visualizations across all combinations of variables hoping that something will "jump out". Create visualizations to help you understand the situation, or to help you answer a particular question. Do not create them arbitrarily or you will spend the entire weekend doing nothing else. Keep track of the most interesting visualizations as the weekend evolves, and use the most insightful visualizations as part of your narrative at the end of the weekend.
7. Make a promise that your team will present their findings regardless of the success of your analysis. You're here for the experience, and it's perfectly acceptable to say “Here was our idea, but it didn't work out because…” The worst that can happen is you'll get some good advice, and (no matter what outcome) you will have a great story to tell in your job interview when they ask “tell us about how you handle failure.”
8. These data are "real", and real data sometimes present unexpected issues.

Prizes

- **Best Insight:** This is given to the team that poses and answers (or partially answers) a question or questions that directly relate to the challenge. It is important that the analyses be described in sufficient detail for judges to understand whether the conclusions are appropriate. It is far, far better to have cautious conclusions that match the analysis (and the scope of the data) then to make a bold claim that you cannot support, no matter how interesting.

- **Best Visualization:** Strong visualizations speak for themselves, almost. Some guidance is almost always needed, and that's fine, but what "sells" a visualization is that it answers a question and provides perhaps some unexpected insight.

- **Best Use of External Data:** While the datasets provided have over 150 variables each, and you might create your own additional variables, your analysis might still benefit from additional context. There are a variety of reliable data sets available from demographic, political, health, and medical sources. This prize is awarded to teams that bring insight to the challenge by expanding the context with this additional data set.