# **Introduction**

Exploring data sets is a way to view statistics and trends that can be useful in creating data visualizations. There is so much data on the internet, that you can use the data that is available to use instead of reinventing “the wheel” by collecting your own data. There are many repositories now on the internet that hold data that you collect or that someone else collects. By accessing the data around you can gain an important perspective on what data you can find online to support your data analysis. For this data curation project, I will focus on one data set from a repository. After reviewing several repositories such as Figshare, Zendodo, Kaggle, and Data.gov, I found an interesting data set about mushrooms from Kaggle entitled “Mushroom Classification Safe to Eat or Deadly Poison. Being new to curating data sets, I found this data set informative, clear, concise, user friendly, which made it easy for a beginner to read, comprehend, and locate the information that I needed. This data set not only benefits the scientific community, but it is a useful data set in distinguishing between edible and poisonous mushrooms for the average “shrooming” individual.

# **Data Information**

The Mushroom Classification data set contains hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family from The Audubon Society Field Guide to North American Mushrooms (1981). The mushrooms were further broken down into two categories: edible and non-edible (definitely poisonous) including those of unknown edibility and therefore, not recommended. The data set, Mushroom Classification Safe to Eat or Deadly Poison, is owned and maintained by UCI (University of California -Irving) Center of Machine Learning. The most current version of this data set was created December 1, 2016. It was originally donated to UCI Machine Learning by Jef Schlimmer on April 27, 1987. Although this data set was created over thirty years ago, it is still used often today. On Kaggle, it has received over 900 kernels, which shows how much interest there is about this data set. The Mushroom Classification data set is in public domain (CCO), which means there is no copyright. Anyone can copy, modify, distribute, and use the information even for commercial purposes without asking permission for use. The data set file is a zip file containing one file , which can be opened by using the notepad app or the Excel program on the computer. There are 8,124 rows describing the 22 different attributes of mushrooms. Each attribute is further divided into various descriptive words. For example, the attribute describing the cap-shape is further broken down into bell=b, conical=c, convex=x, flat=f, knobbed-k, and sunken=s.

# **Metadata Information**

Datasets that use a common metadata standard helps to develop interoperability. This also makes discovery of existing data easier. There are several reasons why metadata standards should be used in datasets. Datasets that use metadata standards can:

* build aggregators to run the same search over
* put the data in context so it easier to read
* help see relationships in the data
* save the next person time and effort without rebuilding the datasets
* allow others to build on and integrate the data

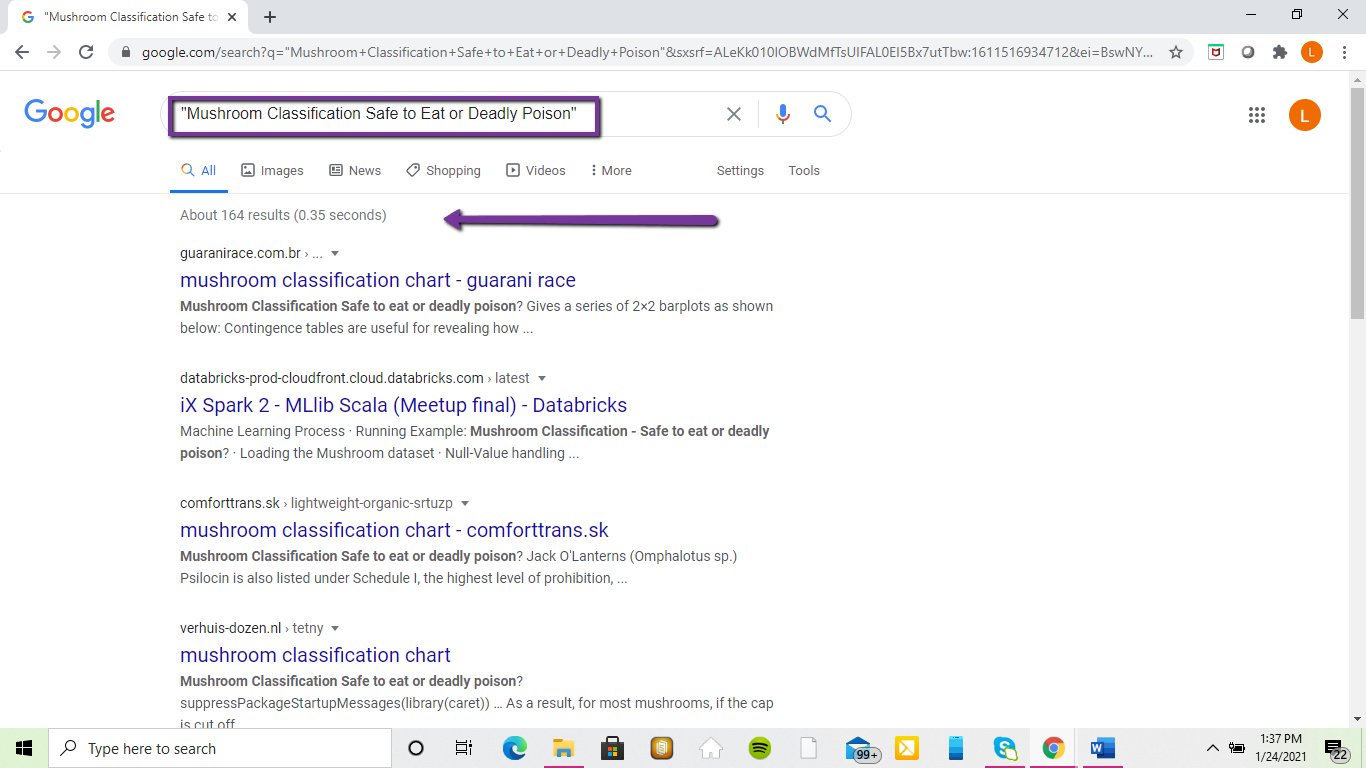
The Mushroom Classification data set is a machine learning data set which uses a structured metadata standard called a tabular data package. It uses the extension csv (comma-separated values), for example in the data set it is labeled as mushrooms.csv. This tabular format uses rows and columns, which corresponds to the record and the features. It also uses the kernel method, which means in machine learning that it uses algorithms for pattern analysis. This is used to study and find the relationships in correlation, classification, and ranking in the data set. This means that the metadata is included in the data files.

**Improving the Current Data and Metadata**

The data set is easy to understand and offers essential metadata. It contains file descriptions, tags, subtitles, and has a public kernel which is fairly active. The notebook feature that allows people to share their work with others. It also offers a discussion thread that allows people to ask questions and make comments or suggestions. Some suggestions for further improvement are to enhance the visualization and readability. This data set would benefit with better data visualizations. The presentation of the data could also be presented in the form of bar graphs, pie graphs, and scatter graphs, just to name a few data visualization enhancements. Showing the pictures of the mushrooms themselves will aid in visualizing the different species of mushrooms and more easily understand the data better, for example, pictures of the various types of mushroom caps. Visualization of the poisonous and non-poisonous mushrooms would help increase recognition of the mushrooms if the data is used for the “shrooming” individual.

**Publications that used the Data Set**

Kaggle does not show or list any publications that have cited this data set. The owners (University of California, Irving Machine Learning Repository) of the data set, Mushroom Classification Safe to Eat or Deadly Poison, used rexa.info to show how many and what publications cited this data set in their research. Forty-seven papers have cited this data set in their publications. In doing a google scholar advance search, I typed in the exact phrase box the name of the data set. The search resulted in one source (Sabrina Ingrid Davita) used the data set at rbubs.com on December 1, 2019. However, when I did a google search I typed “Mushroom Classification Safe to Eat or Deadly Poison” it yielded 164 results. See the screen shot below.



## **Conclusion**

Repositories and data sets are good resources for data collection. By using data sets that are already available saves time and allows for people to build on other concepts and information. Data sets can foster creativity and data sharing using the internet through online repositories that hold the data sets. In the past, without the internet this sharing was not possible. Data sharing allows people to create useful data visualizations that can benefit many different fields such as medicine, education, agriculture, statistics, and more.

**References**

DataOne Best Practices Primer (review section 5.4- Describe: Data Documentation); Available <http://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf>

Frye, Curt. (2019) Learning Public Data Sets <https://www.lynda.com/Excel-tutorials/Learning-Public-Data-Sets/5034173-2.html>

Introduction to Machine Learning Tutorials <https://www.educba.com/machine-learning-datasets/>

Mushroom Classification Safe to Eat or Deadly Poison Data Set <https://www.kaggle.com/uciml/mushroom-classification> also located on <https://archive.ics.uci.edu/ml/datasets/mushroom>

Research Data Alliance (RDA) Metadata Directory, <http://rd-alliance.github.io/metadata-directory/>