Name: Project and Competition Basics

Description: A set of solutions for common problems scientists encounter before being able to begin EDA and model building.

Components:

1) How to Successfully add large data sets to Google Drive and use them in Google Colab.

https://www.aboutdatablog.com/post/how-to-successfully-add-large-data-sets-to-google-drive-and-use-them-in-google-colab

“””

This article shows a person how to load data from Google Drive, reliably and quickly; even with very large datasets.
The most common problem facing data scientists working with large datasets is how to explore the data without access to superior hardware.
Google Colab is the most cost efficient solution, but with very large datasets needing to load the data each instance can be prohibitive in time consumption. This problem becomes exasperated with a poor internet connection.
”””

[This first problem that every project involving large files faces is how to even access these files. It is a challenge that unless overcome, prevents any further exploring in a competition.
Drone Classification: Three 2gb+ files
Streamflow Prediction: Five 3gb+ files
Geological Mapping: “Huge” files (two or more files of 2gb or larger in size)
Mineral Prospecting: One 3gb+ file that is intricately intercorrelated, many categorical variables
Freshwater Management: Two+ 300mb+ files

Any AISC competition in the future with large files will create this identical hurdle for everyone without access to a powerful computer.]

{Fork: The Google Colab Recipe, by Ammar Khan}

2) Tips & Tricks for Working with Large Datasets

https://www.kaggle.com/code/frankherfert/tips-tricks-for-working-with-large-datasets/notebook

“””

This article delves into a diversity of easy to implement tips to reduce the memory consumption of a file without chunking the file or eliminating data.

When manipulating a dataset in Google Colab, a diametric hurdle is the 12gb of memory. Once maxed out, the instance stops and restarts (any model that was currently running will need to also be restarted). Reducing the memory consumption of a large file can be critical.

For exceptionally large files that would max out 12gb simply being loaded to pandas before being modified through the suggestions in this article; these modifications would need to be implemented on a host system first.
“”””

+ How to observe memory consumption of a data frame, as well as each of its columns.
+ Convert dates to datetime data types
+ Converting text columns with repeating values (e.g. Cities or Employees) into a categorical dictionary, as well as a pervasive list ofconverting other data types and layouts into category and numerical types
+ Saving objects as pickle-files for faster loading
+ Manually controlling Python’s garbage collector function

[With the Mineral Prospecting dataset, loading the dataset to pandas caused memory errors in Colab. The file needed to be modified on my host system (bereft of a Nvidia graphics card) first.]

3) Reducing Pandas Memory Usage for Large Datasets #1: Lossless compression

https://pythonspeed.com/articles/pandas-load-less-data/

“”””
Reducing the memory load of a dataset in Pandas without losing any data is the powerful first step for working on multi-gigabyte projects.
“”””

+ Only loading specific columns
+ Converting to lower consumption numerical data types
+ Converting to lower consumption categorical data types
+ Sparse columns

{Optional: Reducing Pandas memory usage #2: lossy compression, Reducing Pandas memory usage #3: Reading in chunks, Fast subsets of large datasets with Pandas and SQLite, Loading SQL data into Pandas without running out of memory, Saving memory with Pandas 1.3’s new string dtype, From chunking to parallelism: faster Pandas with Dask, Reducing NumPy memory usage with lossless compression, NumPy views: saving memory, leaking memory, and subtle bugs}

Pandas:
Visualizing Many Columns: https://www.kaggle.com/code/jeongbinpark/tps-jun-how-to-visualization-many-columns
Pandas: An Essential Exploration, With Brandon Rhodes : https://speakerdeck.com/pycon2015 and https://github.com/PyCon/2015-slides

Kaggle- House Prices: A More Thorough Walkthrough

: https://www.youtube.com/watch?v=_-UCcuB8nbw

SQL:
SQL Tutorial - Full Database Course of Beginners: : https://www.youtube.com/watch?v=HXV3zeQKqGY&t=205s

Deep Learning:

Pytorch Tutorial:

Tensorflow: