

Asset Pricing Dataset Documentation

Empirical Finance & Machine Learning

1 Dataset Overview

This dataset contains monthly equity data for U.S. publicly traded firms, processed for empirical asset pricing and machine learning applications. The data spans from *January 1957 to December 2021*.

The raw data originates from the *Center for Research in Security Prices (CRSP)* and *Compustat*, accessed via *Wharton Research Data Services (WRDS)*. It has been processed to align with the structure used in recent academic literature, specifically Gu, Kelly, and Xiu (2020) [1].

2 Variable Descriptions

The dataset is structured as a panel. The key variables are defined below.

2.1 Identifiers

- ***permno***: This stands for *PERManent NOmber*. It is a unique numeric identifier assigned by CRSP to a security. Unlike ticker symbols (e.g., AAPL, TSLA), which can change over time or be reused by different companies, a ***permno*** never changes throughout the history of the firm.
- ***month***: The timestamp for the observation, representing the end of the month for ***ret_excess*** and the beginning of the month for all other variables.

2.2 Target Variable

- ***ret_excess***: The monthly excess return of the stock ($R_{i,t} - R_{f,t}$). This is the target variable for prediction models.

2.3 Macroeconomic Predictors (***macro_****)

These variables capture the aggregate state of the economy and are identical for all firms in a given month. They are widely used to predict time-varying risk premiums.

- ***Definition***: Variables that reflect broad economic conditions such as inflation, liquidity, and market valuation levels.
- ***Source***: These variables are derived from the influential work of *Welch and Goyal (2008)* [3], who compiled a comprehensive list of variables that have been shown to predict the equity premium.
- ***Examples***:
 - ***macro_dp***: Dividend-Price Ratio.
 - ***macro_tblr***: Treasury Bill Rate (proxy for the risk-free rate).
 - ***macro_tms***: Term Spread (difference between long-term and short-term government bond yields).

2.4 Firm Characteristics (characteristic_*)

This dataset includes *94 distinct firm-specific signals* (features).

- *Source:* These characteristics are based on the extensive inventory of anomalies documented by *Green, Hand, and Zhang (2017)* [2]. They were specifically selected and standardized for machine learning comparisons by *Gu, Kelly, and Xiu (2020)* [1] to represent a broad “factor zoo”.
- *Examples:*
 - `characteristic_mom12m`: Momentum (cumulative returns over the past 12 months).
 - `characteristic_mvel1`: Log of Market Equity (Lagged Size).
 - `characteristic_bm`: Book-to-Market Ratio.
- *Note:* These features are rank-transformed to the interval $[-1, 1]$ to handle outliers and ensure stationarity.

2.5 Industry One-hot Encoded Variables (sic2_*)

These variables control for industry-specific fixed effects.

- *What is an Industry?* We classify firms based on their *Standard Industrial Classification (SIC)* codes. This dataset uses the first two digits of the SIC code to define broad sectors (e.g., Manufacturing, Finance, Technology).
 - If a firm belongs to industry group 35, the variable `sic2_35` will be *1*, and all other industry variables for that firm will be *0*.

3 Working with the Data in Google Colab

The data is provided in Parquet format. When using Google Colab, it is inefficient to upload large files to the runtime session every time, as files are deleted when the runtime disconnects.

Recommended Workflow: Upload the data folder to your Google Drive once, and then “mount” your Drive to Colab.

```
1 from google.colab import drive
2 import pandas as pd
3 import pyarrow.parquet as pa
4
5 # 1. Mount Google Drive
6 # You will be prompted to authorize access to your Drive
7 drive.mount('/content/drive')
8
9 # 2. Load Data
10 # Replace 'MyDataFolder' with the actual path in your Google Drive
11 file_path = '/content/drive/MyDrive/MyDataFolder/202112.parquet'
12
13 table = pa.read_table(file_path)
14 df = table.to_pandas()
15
16 print(df.head())
```

4 Accessing Raw Source Files

If you wish to inspect the raw data inputs used to generate the Parquet files, they are available below.

- *datashare.csv*: Contains the raw firm characteristics [2].
- *tidy_finance_python.sqlite*: Contains monthly returns and macro data.
- *About SQLite*: SQLite is a lightweight, serverless database engine. Unlike traditional databases (like MySQL), the entire database is stored as a single file on the disk, making it easy to share and read using Python.

4.1 Code to Load Raw Files in Colab

You can download the files directly to the Colab runtime and read them as follows:

```
1 import pandas as pd
2 import sqlite3
3
4 # 1. Download files (using direct download links)
5 !wget -O datashare.csv "https://www.dropbox.com/scl/fi/7
   xoe7286nl451p39s7rvz/datashare.csv?rlkey=cs5ca0zmdgjbb1tgwun3ie3e6&e
   =1&dl=1"
6 !wget -O tidy_finance.sqlite "https://www.dropbox.com/scl/fi/
   e70qgc94j1uyby6aqz82z/tidy_finance_python.sqlite?rlkey=6
   nnnyxich0f21joq4ag17aq7&st=0og9rayn&dl=1"
7
8 # 2. Read the CSV (Firm Characteristics)
9 # Reading only the first 1000 rows as a sample
10 df_features = pd.read_csv('datashare.csv', nrows=1000)
11
12 # 3. Read from SQLite (CRSP & Macro Data)
13 con = sqlite3.connect('tidy_finance.sqlite')
14 # Querying the CRSP monthly table
15 df_crsp = pd.read_sql_query("SELECT * FROM crsp_monthly LIMIT 5", con)
16 con.close()
17
18 print("Features Shape:", df_features.shape)
19 print("CRSP Data Shape:", df_crsp.shape)
```

5 Recommended Reading

Students are encouraged to read and consider the following papers to deepen their understanding of machine learning applications in asset pricing and tabular data.

- *On Deep Learning for Tabular Data*: Ye et al. (2024) provide a closer look at deep learning methods specifically tailored for tabular datasets [4].
- *Recent Advances*: Students should also consider the recent working paper available on SSRN regarding advanced modeling techniques [5].

Acknowledgments

Pdatasetarts of the code generation process were assisted by an AI language model.

References

- [1] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223-2273.
- [2] Green, J., Hand, J. R., & Zhang, X. F. (2017). The characteristics that provide independent information about average U.S. monthly stock returns. *The Review of Financial Studies*, 30(12), 4389-4436.
- [3] Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4), 1455-1508.
- [4] Ye, H. J., Liu, S. Y., Cai, H. R., Zhou, Q. L., & Zhan, D. C. (2024). A Closer Look at Deep Learning Methods on Tabular Datasets. *arXiv preprint arXiv:2407.00956*. Available at <https://arxiv.org/abs/2407.00956>
- [5] Liu, Y., Luo, Y., Wang, Z., & Zhang, X. (2026). Uncertainty-Adjusted Sorting for Asset Pricing with Machine Learning. *SSRN Working Paper No. 6013574*. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6013574.