Extraction, Transformation, and Load Technical Report

Tariq Attarwala, Sai Reddy, and Veronica Valencia

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Summary

We seek to organize Standard & Poor(S&P) 500 and macro-economic data. The organized data source will provide us the opportunity to understand the impact of 1.Inflation(CPI), 2.Employment(U.R), 3.Housing(HPI), 4.Consumer Spending, and 5.Consumer Confidence on S&P 500 between 2018-2020.

1.2 Scope

The macro-economic data will be extracted from FRED (Federal Reserve Economic Data) and the S&P 500 from yahoo finance. An ERD diagram that depicts the data lineage along with the primary and foreign keys for each table will be provided. A data dictionary that provides variable description and data type will also be included as part of the project. No analysis will be done on the data extracted.

1.3 Technologies and resource contributions

Extract - Tariq

- Extract the Macro-economic variables from FRED
- Extract the S&P data from Yahoo finance

Transform- Veronica

- Read in and transform the extracted data into the required format using python.
- Create an ERD diagram that shows table relationships, data types and Keys using database diagrams.

Load- Sai

- Load the transformed data in postgres DB using python.
- Check the data quality, accuracy and integrity of the tables uploaded using postgresql.

1.4 Definitions, Acronyms and Abbreviations

ETL - Extract Transform Load
FRED - Federal Reserve Economic Data
S&P 500 - Standard & Poor 500

2. ETL DETAILS

This section outlines a more detailed description of the processes utilized/proposed to achieve the objectives of this initiative.

2.1 Data Import/Extract Sources and Method

The data pulled from this project is from Yahoo Finance and from FRED. Below you will find the links for the data, which has been downloaded as .csv files

S&P500 data:
https://finance.yahoo.com/quote/%5EGSPC/history?period1=1553558400&period2=1590451200&interval=1d&filter=history&frequency=1d

Inflation Data (Consumer Price Index):
https://fred.stlouisfed.org/series/CPIAUCSL

Unemployment Data:
https://fred.stlouisfed.org/series/UNRATE

Consumer confidence (sentiment):
https://fred.stlouisfed.org/series/UMCSENT

Housing Data:
https://fred.stlouisfed.org/series/CSUSHPISA

Consumer Spending Data:
https://fred.stlouisfed.org/series/PCE

This section provides information about the data and its source. For example, API names and URLs, key parameters available and its subset which will be preserved (loaded). Data extraction protocols (API, FTP, Web scraping etc.), any permissions required to access the said extraction dataset and any restriction placed on the usage and distribution of the acquired dataset.

2.2 Data Acquisition

The macro-economic data will be extracted from FRED in csv format and the S&P 500 data from Yahoo finance. The macroeconomic data and the S&P 500 data are aggregated by month. Both the tables will be refreshed monthly and there is 1 month lag between when the latest data is published and is uploaded to the tables.

2.3 Data Transform

The S&P 500 data extracted from yahoo finance is aggregated by month, by taking the average of S&P 500 closing price. The Year-Month variable in both these tables will serve as primary and foreign key, which can be used to merge the tables for analysis.

2.4 Data Integrity

Missing value, Frequency, summary statistics and distribution reports will be generated every month to insure data quality and integrity. Analysis will be performed on any outliers, unexpected data entries and addressed immediately.

2.5 Data Refresh Frequency

The tables will be refreshed on the last saturday of every month. During the duration of the upload, users will not be able access the data. An email will be sent out to all authorized users prior to and after the data loading process.

2.6 Data Security

Steps will be taken to ensure that the data does not contain any personal identifying information(PII), any PII found will be anonymized.

2.7 Data Loading and Availability

The ERD diagram in the appendix along with the data dictionary provides the necessary information to extract the data. The client can retrieve the data by connecting to the economy_db in postgres.

3. Data Quality

Sample records will be extracted from all the tables to check for data accuracy, after the data has been loaded. Aggregate analysis on the counts will be performed on all the constituent tables to ensure that the data load was successful. Missing value report and summary statistics will be generated on all the columns to validate the ETL process.

4. User Guide

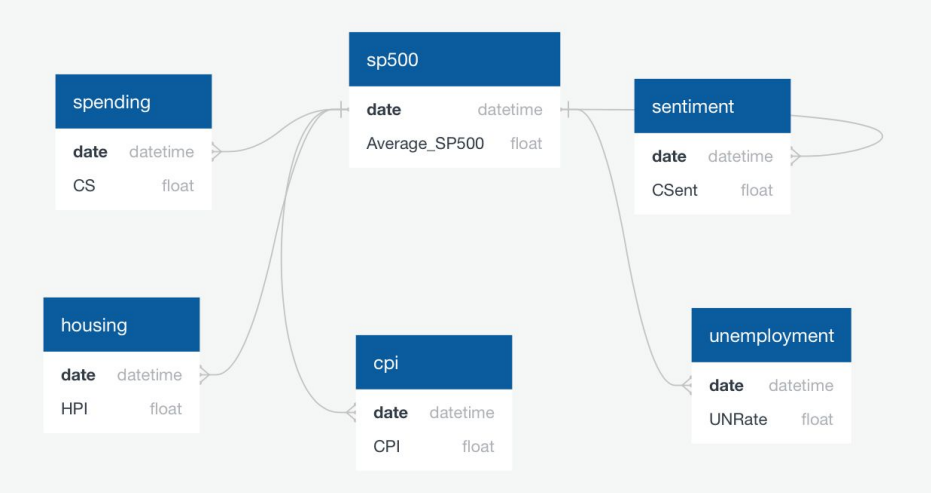
4.1 Data Dictionary

The table below provides the column names, variable description and data type for all the columns involved in the ETL process.

Variable Name	Description	Data Type
Date (PK, FK)	Date filed	date
Average_SP500	Average of S&P 500	float
HPI	Housing Price Index	float
CPI	Consumer Price Inflation	float
CS	Consumer Spending	float
UNRate	Unemployment Rate	float
CSent	Consumer Sentiment	float

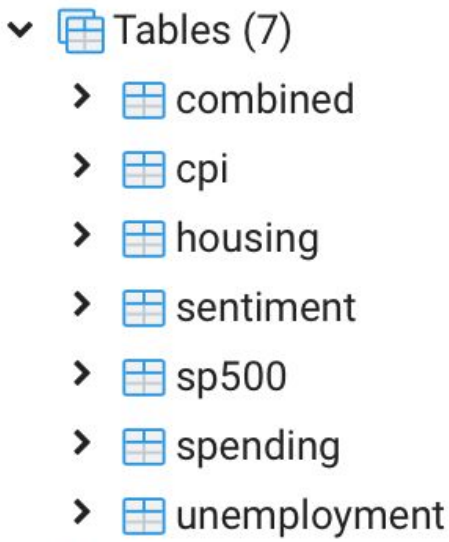
4.2 ERD Diagram

The ERD diagram below shows all the tables that were part of the economy_db, the diagram also identifies the Primary Key (PK) and Foreign Key (FK) corresponding to each table and their relationship with other tables.



4.3 Data Schema

The schema below shows all the tables that were part of the economy_db.



4.4 Sample data

The table below shows the same data for the combined table .

Data Output | Explain | Messages | Notifications

	date text	Average_SP500 double precision	HPI double precision	CPI double precision	CS double precision	UNRate double precision	CSent double precision
1	2019-0...	2818.42	207.566	254.095	14354.6	3.8	98.4
2	2019-0...	2903.8	208.218	254.94299999999998	14452.5	3.6	97.2
3	2019-0...	2854.71	208.693	255.167	14516	3.6	100
4	2019-0...	2890.17	209.032	255.40200000000002	14565	3.7	98.2
5	2019-0...	2996.11	209.442	256.087	14644.4	3.7	98.4
6	2019-0...	2897.45	210.204	256.29400000000004	14682.4	3.7	89.8
7	2019-0...	2982.16	210.955	256.593	14707.8	3.5	93.2
8	2019-1...	2977.68	211.774	257.22900000000004	14745.4	3.6	95.5
9	2019-1...	3104.9	212.663	257.824	14792.5	3.5	96.8
10	2019-1...	3176.75	213.601	258.444	14847.1	3.5	99.3
11	2020-0...	3278.2	214.56	258.82	14913.7	3.6	99.8