**Dataset Name:**

Global Food Prices Database

**Dataset Description:**The Global Food Prices Database has data on food prices (e.g., beans, rice, fish, and sugar) for 76 countries and some 1,500 markets. The data goes back as far as 1992 for a few countries, although most of the price trends start in 2000-2002.

**Data Set analysis:**

**Data Dictionary, meaning, and analysis of every column:**

|  |  |
| --- | --- |
| ADM0\_ID | ID for the particular administrative boundary – this is actually the ID of the country. There are 21 countries represented. 30% of the data belongs to the markets in Republic of Congo, followed by approximately 10% in Burkina Faso and Colombia, 7-8% to Armenia and Cameroon and rest to the other countries. All these countries are underdeveloped countries. |
| ADM0\_Name | Name of the country; same as above |
| ADM1\_ID | ID for the particular administrative boundary – actually the ID of the province; There are a total of 137 provinces belonging to 21 countries. In a few cases i.e. 795 cases, the province ID/Name is missing |
| ADM1\_Name | Name of the province or state corresponding to the above |
| Market\_ID | ID for this particular market. (There are a total of 220 markets) |
| Market\_Name | Name of the market (There are only 217 unique market names; it means that the same market names have been assigned to more than 2 markets. This can easily be fixed by looking up other market IDs and their respective market names and assigning the correct market names to those markets that have the names assigned wrongly) |
| Commodity\_ID | ID identifying the specific commodity. A total of 139 different commodities have their pricing data in this dataset. |
| Commodity\_Name | Name of the commodity; same as above |
| Currency\_ID | ID for the currency in which trading took place; The prices are in 16 different currencies. |
| Currency\_Name | Name of the currency; The database contains pricing data in 16 different currencies. |
| Price\_Type\_ID | Price type id – wholesale or retail price; 82% of the prices are retail prices and rest 18% are wholesale prices. |
| Price\_Type\_Name | Price type name – whole sale or retail price; same as above. |
| UM\_ID | ID corresponding to the unit of measurement; there are 21 unique units of measurement ranging from KGs to galons, units, grams, pounds etc. |
| UM\_Name | Name for the above i.e. Name in string such as KG/90KG/Units etc |
| MP\_Month | Month in which this price was recorded; The prices records are almost equally distributed with 8 to 9 % records pertaining to every month of the year. |
| MP\_Year | Year in which the price was recorded; The prices are from Jan 1992 to Jan 2017 and 75% of the records pertain to the current decade from the year Jan 2011 to Jan 2017 and they are almost equally distributed over all the years. So, this data is relatively recent and may be quite useful. |
| MP\_Price | Market Price for the commodity in this month; We have this data available for all the records. |
| MP\_Commodity\_Source | Source from which this data was received; The source of data is very genuine and come from organizations such as Food and agricultural organization(FAO) or World Food Program (FWP) of United nations, Ministry of Agriculture of different countries etc. |

On analyzing the data sets, esp. the country (all of the countries are either undeveloped or underdeveloped countries of Asia, Africa, and South America), source of data (such as FAO/FWP/ministry of Agriculture etc.), it has been observed that there is very strong possibility that this data pertains to the prices at which the above organizations procured the food items in different countries/provinces etc. for overcoming hunger alleviation of the deprived sections of these countries. It can be safely said that this data doesn’t belong to the commodity exchanges because of a no. of reasons such as source of the price data, availability of only monthly prices etc. If it was trading data, we would have at least daily prices and ideally hourly prices.

**Data Quality Issues:**

**Missing Values:** In 795 cases records, the id/name of the province to which the commodity market price belongs to is missing. However, we don’t need to remove these records, because country id and name are present and since these countries are small, we can assume the other provinces of the same country to be the provinces in case of those records where the province details are missing. Other than the above fields, data is in healthy shape and all other variables don’t have any missing value.

**Errors:** As discussed in the data-dictionary, if we analyze the fields Market-ID and Market-name, we observe that there are 220 different Market IDs but only 217 unique market names. Please note that there is a one to one mapping between market ID and market name. This implies that 3 markets have been assigned wrong names. However, the lack of market name doesn’t significantly affect the user of the data because the price would be approximately remain the same in the same province or country. As such we don’t really need to drop/discard such rows.

**User requirements: -**

Organizations such as FAO/FWP/Ministry of agriculture/Ministry of Human welfare/ NGOs like Bill & Melinda Gates foundation may be interested in this data so that they can analyze the past prices of different food items across different geographies and in different currencies, thus enabling them to procure the food items at the cheapest price possible in future.

**Pros and Cons of Various Modelling schemas:**

**Star schema (advantages):**

* Easy to understand and visualize for the business users
* Data can be analyzed and visualized easily from different dimensions and perspectives
* Very fast query performance, because minimal joining takes place; it’s a flat representation of data and many values often get repeated

**Star schema (disadvantages):**

* Redundancy of data – no 3NF
* Because of redundancy of data, it may be prone to errors while making changes.
* If changes are to be made, the same changes have to be made everywhere.
* More storage requirements

**Snow-flake schema (advantages):**

* Generally 3NF and hence less redundancy and hence less storage requirements
* Easy to maintain
* More flexibility

**Snow-flake schema (disadvantages):**

* May be difficult to visualize for business users from business perspective
* Less efficient in terms of query performance because lots of joining may take place

**Data Vault (advantages):**

* Closer to the business organization
* Easier to maintain
* More flexible – easier to create hubs/satellites/links when a new business unit is added or removed
* Less redundancy

**Data Vault (disadvantages):**

* Difficult to visualize from analytical point of view, for the business users
* Query performance is not good because a lot of joins may take place

**Data Warehouse Approach Selection:**

We have followed a multi-dimensional and hybrid approach – Star and Snowflake, because we see that the data pertains to the food procurement by companies such as UN-FPO or similar such organizations that may need to analyze the food items procurement wholesale/retail prices from different perspectives such as month/year/place/country/market/currency. The users of such kind of databases are not generally very tech savvy and would also need a very good query performance. As such a hybrid model consisting of primarily Star based model with a fact table and various dimension tables in respect to various dimensions from which the analysis can be done. Nevertheless the Star schema would be briefly supported by snowflake model too.

There would be one fact table and several dimension tables.

The fact table would contain the prices of different food commodities in poor countries in different currencies. There would be different dimensions of the food prices which may be markets, countries, provinces, commodities, retail/wholesale price types etc. Some of the dimension tables may further be supported by other tables, thus giving the whole data warehouse approach as a hybrid of star and snowflake approach.

**Data Warehouse Design:**

The MWB file is attached. However, following is a pictorial snapshot of the same.

