Commodify

CIS 550 Final Project Proposal

Team Members

Name	Email	GitHub
Iris Tiong	iristyx@seas.upenn.edu	iristyx
Francis Featherby	ffrancis@seas.upenn.edu	ffeatherby
Peter Brice	pbrice@seas.upenn.edu	briceybrit
Steven Brooks	sbr@seas.upenn.edu	stevegbrooks

Description of project

We will build a database containing supply and demand information for various commodities (metals, other minerals, oil and gas, agricultural commodities etc.), containing historical information as well as projections, and we will have a feature that allows the user to input their own estimates for various values - for US crude oil production in 2021, for example - to see the implications for supply and demand. As well as a search page we will have a 'dashboard' page, potentially customisable, allowing the user to see the most recent data of relevance to them, which might be the most recent supply and demand data and/or the most recent prices for given commodities.

We plan to use charts and maps to visualize the data.

Our project is an innovation because existing resources deal mostly with individual sectors like energy or agriculture, but not the the whole commodities space, and they are read only. Our project provides a useful resource for trade houses, banks and hedge funds.

Datasets

1. Current and historical agricultural commodities data from the USDA Foreign Agricultural Service

- Description: a dataset containing commodities and their prices around the world, including trading value at the beginning and end of each month.
- Size: ~200 MB with ~2 million rows and 12 features.
- A commodity is identified by "Commodity_Code", and also has location, time, and various attributes attached to its value.
- A quick peek at the first 5 rows of the data shows the following (column names changed for brevity; not all columns shown):

ccode	comm	cntry	mrkt_yr	year	month	attr_id	attr	unit_id	unit_desc	val
577400	Almonds	AF	2010	2018	10	20	Beginning Stocks	21	(MT)	0
577400	Almonds	AF	2010	2018	10	125	Domestic Consumption	21	(MT)	0
577400	Almonds	AF	2010	2018	10	176	Ending Stocks	21	(MT)	0
577400	Almonds	AF	2010	2018	10	88	Exports	21	(MT)	0
577400	Almonds	AF	2010	2018	10	57	Imports	21	(MT)	0

- Number of unique values for each column (original column names):

col_name	unique
Commodity_Code	63
Commodity_Description	63
Country_Code	212
Country_Name	213
Market_Year	62
Calendar_Year	62
Month	13
Attribute_ID	71
Attribute_Description	71
Unit_ID	11
Unit_Description	11
Value	44725

- Summary statistics for numeric columns:

Market_Year	Calendar_Year	Unit_ID	Value
Min. :1960	Min. :1959	Min. : 2.000	Min. : -9510
1st Qu.:1979	1st Qu.:2006	1st Qu.: 8.000	1st Qu.: 0
Median :1992	Median :2006	Median : 8.000	Median : 15
Mean :1992	Mean :2006	Mean : 9.833	Mean : 11937
3rd Qu.:2007	3rd Qu.:2014	3rd Qu.: 8.000	3rd Qu.: 212
Max. :2021	Max. :2021	Max. :29.000	Max. :42528700

2. Current and historical energy supply and demand data from the Energy Information Agency, part of the United States Department of Energy

- Description: data sets released weekly and monthly containing thousands of data points on US and global energy production and consumption.
- Size: the data are provided in various packages and formats, some overlapping, but there are at least hundreds of features per week in data stretching back decades, i.e. thousands of rows.
- The data are mostly time series. They show various aspects of supply and demand, e.g. production, consumption, inventories etc., for a given geographical region, such as a state or country, on a sequence of dates of in a sequence of periods.
- Below are some example lines from a CSV file dealing with US crude oil production in the lower 48 United States by month in 2020 in thousands of barrels (column names changed for brevity).

Date	U.S. Prod	East Coast Prod	Florida Prod	New York Prod
2020-06-15	313264	1967	69	23
2020-07-15	340152	1968	122	23
2020-08-15	328099	2189	122	23
2020-09-15	326114	2309	108	23
2020-10-15	323387	2359	113	23
2020-11-15	333721	2180	117	23

- Summary statistics:

Date	US Prod	East Cost Prod	Florida Prod	New York Prod
Min. :1981-01-15 00:00:00	Min. :119208	Min. : 399.0	Min. : 35.0	Min. : 9.00
1st Qu.:1990-12-30 12:00:00	1st Qu.:175752	1st Qu.: 649.5	1st Qu.: 174.5	1st Qu.:19.00
Median :2000-12-15 00:00:00	Median :212585	Median : 861.0	Median : 370.0	Median :28.00
Mean :2000-12-14 11:43:27	Mean :223392	Mean :1171.7	Mean : 511.4	Mean :33.37
3rd Qu.:2010-11-30 00:00:00	3rd Qu.:263536	3rd Qu.:1518.0	3rd Qu.: 542.5	3rd Qu.:35.50
Max. :2020-11-15 00:00:00	Max. :396865	Max. :4243.0	Max. :3606.0	Max. :96.00

Data Queries

1. Top almond producers around the world in countries that consume the highest amount of oil.
2. Ending market prices for March 2020 of the top commodities produced by metric ton of that month.
3. Projected US corn ending stocks for 2021 based on the user's estimate for this year's harvest.
4. Which American states were amongs the top ten producers of both agricultural commodites and energy products?
5. (If weather data is incorporated) What is the average precipitation in the five states with the highest wheat yield, and for the other states?