Project A9: Crop production analysis

Project repository: https://github.com/katharinakuusk/IDS2021_FAO

- Ulrich Ojoh Dioni
- Katarina Kuusk
- Tsienza Smith Steven

Task 2

Identifying your business goals

Background

Yield of horticultural crops depends on many environmental conditions and factors that partly could be managed by farmers. The Food and Agricultural Organisation acting as a specialized agency of the United Nations that leads international efforts to defeat hunger in the world most especially in sensitive areas and cohorts, collects cash crops production data worldwide with the goal to achieve food security for all and make sure that people have regular access to enough high-quality food to lead active, and healthy lives. This data are made public and can be accessed by featuring organisations concerned with this framework for risk factors assessment, predicting future trends in the agriculture and food industry sector. Understanding these indicators assists decision making of its member states government and producers to insure food security in different countries, equalize rights of producers, regulate international commercial practices, and encourage knowledge transfer from best practices with member states.

Business goals

Mindful of the precedent, we intend to explore the data available in the database while restricted our research goals to apple (*Malus domestica*) which is a common fruit species. Apple is a nutritionally valuable crop distributed in different regions of the world.

We hypothesize that there is a dependence between overall yield of horticultural crops and producer price. We expect to see growth of productivity with the year as a result of technological improvement. There are peculiarities of the crop that may determine this dependence. Apple grows in perennial orchards thus the effect of

unfavorable factors may sustain throughout several years. Apple fruits have a long shelf life, some of the varieties may be stored up to 7 months. Moreover, there have many processing methods that allow conservation of the product in case of yield excess. From this one hand perspective, we would like to observe differences in the behavior of the same variable in other crops. As a final business goal we could to be able to predict future producer price based on the coming yield data from partners and farmers of the countries.

Model prediction will be assessed using a test dataset for a year. A subjective result is conclusion drawn from the study and understanding the direction of future research goal in the topic. FAO has quite comprehensive visualisation tool for the data, basic steps of our analysis could be compared with their system to determine if we were able to relate our findings to their results, or work on some improvement as well as check how right our work is done in the process, provided that machine learning steps will be further performed on a properly prepared data.

Situation Assessment

Inventory of resources

- Dataset: Crop statistics are recorded for 173 products with the objective to comprehensively cover production of all primary crops for all countries and regions in the world. FAOSTAT database provides clear datasets varying from crops production to producer market price by countries and years, yield, yield per area and total yield. Terminology defined in the corresponding section.
- FAOSTAT Terminology: A comprehensive material guide is provided that describes each feature and the data collection aggregates.
- Hardware: The assessment processing of the data will be computationally performed and mobile phone communication to keep in touch with team members.
- Software: “jupyter notebook” will be used as a data analysis tool on a primary research pattern ranging from data exploration, representation, and data mining. Online resources as search engines for search of support materials. “Trello” is used as our managerial board and project brainstorming. “Github” for our online shared repository and a mobile communication tool.

Requirements, assumptions, and constraints

The dataset availability is poised on effective enrollment of each member state given that the target years record of data differs from countries. This directs our observation as some countries began to submit the data to FAOSTAT from different years and also some data entries are based on factual estimations rather than

critical procedural calculations. This could constrains the research on the fact that reliability could tend to present disparities if relation to similar categories.

On this regards, yield and area under the crop data has been recorded from 1961 but the price data is available only 1991. Moreover not all the countries were submitting the indicators simultaneously in 1991, some of them began to provide data less than 10 years ago. Thus, we have to drop these parameters. We couldn't replace them by the median because (maybe only if the gaps are small (1-2 year) and not from the beginning of the data collection) due to the nature of the indicators.

Risks and contingencies

Mindful that there could be potential loss of work materials and resources as we progress in the research, we use "Github" as our shared repository to maintain our work integrity. We organise research objectives in an alternate role play for optimal participation in sub-tasks.

Terminology

Producer prices: are prices received by farmers for primary crops as collected at the point of initial sale (prices paid at the farm-gate). Units: USD/tonne of fresh product/

Area: is total area harvested in the country. Units: hectares

Total yield: in our terminology is the same as "production quantity" in FAO terminology, is a total harvested fresh product in the country. Units: tonnes.

Yield: the same as "yield" in FAO terminology is obtained by the following formula:

Yield = Total Yield / Area

Units: hg / ha 1hg = 0.1 kg = 0.0001 tonnes

Gross index: same as FAO terminology is the indices of agricultural production showing the relative level of yield for each year in comparison with the base period 1999-2001.

Task 3

Data understanding

As we intend to categorize apple production countries by yield, to reveal relations

between yield, yield per area and and producer market in order to predict yield values, we are downloading dataset from FAOSTAT https://www.fao.org/faostat/en/ following the period from 1991 to 2019 since it available . The data on apple that we need to use are those that are expressed in terms of area harvested, production quantity and yield. The data we require here need to cover a comprehensive record of primary crops for all selected countries.

- ***Gathering data***
    - Outline data requirements

Country: All listed UN FAO member states. Since we randomly collect a list of country on the apple production data, this feature is interesting to categorise them in scale of higher producer as to reference those influencing the global market, middle producers for those that would tend to export less and whose production is directed to mainly local consummation and non producer whose dependence are directed to the market availability.

Year: availability of the apple data collection record years for the period of 1991 to 2019. It is ideally interest to meet our research objectives in this range since the FAOSTAT database is missing almost two years data update. Our model prediction should considerable to forecast the upcoming.

Yield: the apple crop tonnage per year as offered to provide meaningful figures on area, yield, production and utilisation. It is ideal for this study case since it is self represented and does not depend on other crops components for its quantification. Also, it does not require to be planted every year. And as per FAOSTAT terminology, it comes directly from the land without undergoing any real processing apart from cleaning.

Yield per area: the apple yield per harvested area of the apple crop for each given year per country. The land aggregate for the apple to is necessary to evaluation if yield in this case is a quantitative or qualitative relation to yield.

Total yield: total yield harvested.

- Verify data availability

From the FAOSTAT portal, we can refine our selection as per our needs. We choose

from Crops and livestock products the list of countries, the production years 1991 to 2019, the apple crop in the primary crops category and the features we are interested in. Our selection field is narrowed at the last two decades since some countries started submitting data not later than 1991.

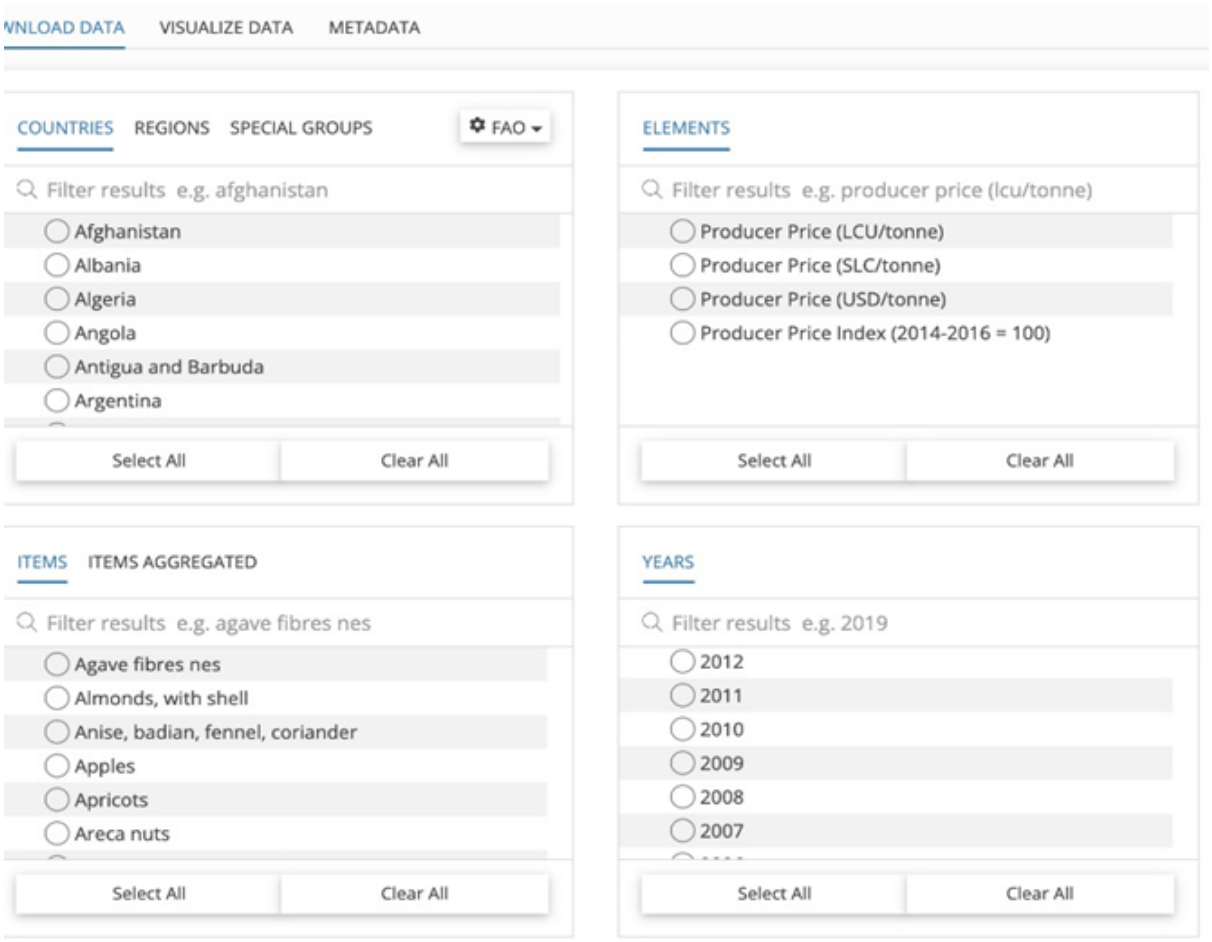

Figure 1: The FAOSTAT database portal will clean datasets of selectable crops production for statistical analysis offering features as countries selection, year periodicity, crops categories and related features as per needs.

·Define selection criteria

The dataset for the selected features are available for download as CSV file template. The data is clean and can be imported in “jupyter notebook” as our data-mining platform. We perform a first hand exploration insight to reveal the potential mining resources we need to carry on the research.

	Country	Year	Area_ha	Price_USD_tonne	Yield_hg_ha	TotalYield_tonnes
31	Albania	1993	2076.0	461.0	48170.0	10000.0
33	Albania	1995	2140.0	323.6	46729.0	10000.0
34	Albania	1996	2100.0	325.4	47619.0	10000.0
35	Albania	1997	2242.0	253.1	50401.0	11300.0
36	Albania	1998	2300.0	248.9	50000.0	11500.0
...	...	...	...	...	...	...
2671	Yemen	2017	2179.0	2052.2	78068.0	17011.0
2672	Yemen	2018	2148.0	1991.4	77775.0	16706.0
2699	Zimbabwe	2015	744.0	1897.0	89516.0	6660.0
2700	Zimbabwe	2016	735.0	1789.0	89116.0	6550.0
2701	Zimbabwe	2017	734.0	1865.0	88747.0	6514.0

[1854 rows x 6 columns]

Figure 2: A preview image of the apple production dataset mainly constituted of six features being country list, years of data collection, producer price, yield per harvested area and total yield in tonnes

·Describing data

The apple production analysis dataset we use is sourced from the FAOSTAT portal, we downloaded four datasets of CSV templates comprising six features of the apple production statistics that is the recorded on yearly basis as apple yield tonnage per harvested area in hectares, the total yield tonnage, the producer price in US Dollar per tonnage, the representing country for the apple product and the year of data record.

·Exploring data

For the selected record from 1991 to 2019, the data have an annual average length of 68 ranging from 35 to 72 over the years. This is an indication that our data have missing values to deal with. And also, this would be interesting to figure out since our first intuition of some countries not submitting their production records at earlier dates does not stand the ground as the dataset length is not progressively distributed to the yearly recording and evolution. We need to find the missing values. We need to find if the values would critically impact our findings like, why a country's apple yield is quantified but does not support tonnage price.

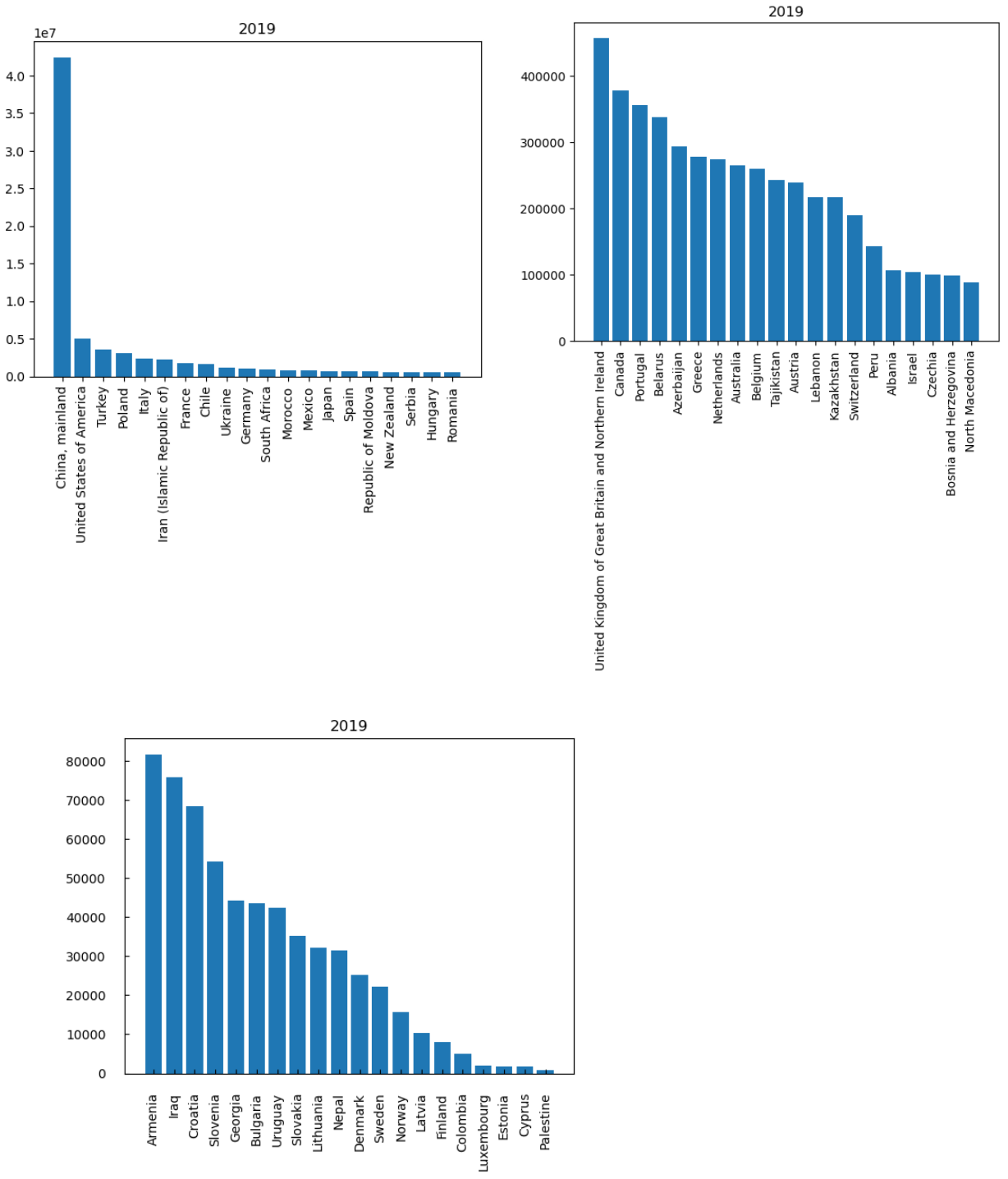

Figure 3. A bar plot for 2019 on total yield in tonnes representing a three potential grouping class from high producers to local consumers.

·Verifying data quality

The first apercu from the data representation highlights the feasibility of our primary goal in this project workflow that is to categorise apple production countries by yield. Further steps will be to deepened our understanding as we put to light the analytic features to support the price per yield variation.

Task 4:

For this project, we work on Crop production datasets provided by FAOSTAT which is a platform that provides free access to food and agriculture data for over 245 countries and territories and covers all FAO regional groupings from 1961 to the most recent year available.

In the agricultural sector, farmers and agribusinesses have to make crucial decisions every day and many of the factors that influence them are highly complex. A key issue for agricultural planning is the accurate estimation of the yield of the many crops involved in the planning.

Data mining techniques are a necessary approach to finding practical and effective solutions to this problem.

After analyzing the dataset, we thought it would be interesting to meet the following goals:

- Categorizing apple producing countries by yield
- to analyse relation between yield, yield/area and producer market price
- to reveal the price variation indices
- try to predict price by yield values

There is concern for a comprehensive apprehension of the data modeling beforehand and the corresponding techniques to implicate.

Task distribution

Ojoh - Data understanding and exploration.

Katarina:

Dataset search - 5 h; business explanation - 3 h , data preparation and assisting data exploration - 5 h; regression modeling and research on the topic and different approaches - 7 h; plots preparation - 3 h; presentation preparation - 7 h

Steven - project planning and data analysis.