Process Book Group 39

June 5:

- Met for the first time as a group
- Made sure everyone had git and python3 installed and created git repository
- Chose dataset: Global Food Prices Database (WFP)
- Had first meeting with TA Houda
- Came up with subquestions:
  - Are there any food prices that are show negative/positive correlation, and is this correlation present throughout the years, or perhaps only in certain period? Can you perhaps detect possible ingredients of a certain other food product?
  - Do countries in similar regions, also show similar price differences? And if differences occur, can you find a potential explanation?
  - What are the correlations in production numbers of crops and the price of certain types of food? Are there correlations in production dominance of certain countries and the price of certain types of food?
- Created this process book
- Created report latex file
- Found database on crop production numbers per country
- Wrote hypotheses to subquestions

June 7

- Have a litte chat about the progress we made each the previous day.
- Write down the problems and prepare for the TA meeting.
- Problems we have:
  - Data cleaning problems, should we convert the prices to dollars or euro’s
  - Missing data in the price data set, we encounter a lot of data which is measured over a short period of time
  - Data problems with the production data set
- Solutions we have for our problems:
  - We decided to make a decision later in the process if we are going to convert the currency. It depends on the outcome of the tests.
  - We decided to try to eliminate data which is measured for a short period of time, still work in progress
  - We reduced the production dataset, we only kept the pure production data and filtered it along the countries we have in the price dataset. Note: the production dataset only contains production data per country per year.
- We have little with the source control on github, but we will be fine

- Matthijs is going to make a start with the eda part of this assignment. Trying to make some pots and do some pioneering work.

- Roel is going to check if the data of the production dataset is complete and determine what is missing. He is also going to find out which production product to link with the price products. (They differ a little bit)

- Tessel is going to finish the data cleaning, so we are ready for next week.

- Jesper is also going to start with the eda part of this assignment, and function as back-up for other problems which will arise.

June 11, 2018 (Monday)

We had a quite long meeting about what we are going to do with the EDA. There were a few problems with the dataset, mainly about the few entries of a few products. We had a little discussion and decided to deal with this after we made some progress with the EDA.

After this we talked a lot about what we wanted to show in the EDA, and what we are searching for. We divided the questions into smaller parts, and came to the following list of what we wanted to show:

- Showing mean food prices per country, in a graphical representation over time.
- Determine non-graphical information about mean food prices per country.
- Combine mean food prices per country in a graphical representation with 2 or more products.
- Combine information about mean food prices for multiple countries from the same region in a plot.
- Find non-graphical data about the production dataset.
- Compare the production dataset with other data and summarize this information in plots.

Finally we decided to adjust our workflow with Github. Every time a new feature to the project is being developed we create a new branch. This helps us to better keep track of what is being done, and reduce the amount of merging conflicts.

From today we each keep track of what we have done every day, and update this in the process book.

Personal logs June 11, 2018

Jesper	9:30 doing research what we need do to in de EDA, create some ideas 13:00 meeting with the group 15:00 update the process book with the information we discussed In the meeting.
Roel	Searched for common products between datasets that have the same exact name, this proved to be insufficient. Wrote code to produce lists of products from both datasets, found the proper common products by hand and put it in an excel sheet.
Tessel	16.00-18.00 Perfomed some data analysis to determine which data the EDA should be based on. Visualizing product frequencies, available years. Located missing entries and searched for strategies to handle these.
Matthijs	Made python functions that generate statistics about a subset of the data. This subset can be produced by a query (i.e. year==2015 & country==” Afghanistan”). Statistics that are generated are mean price, standard deviation, number of entries, number of entries within X times standard deviation (outlier detection). Also made function for a box and whisker plot for price based on query.

June 12, 2018 (Tuesday)

Today we met up early morgen 9:30 for a brief discussion about the progress we made in the past 24 hours. We are very happy with the progress we made, but still have some difficulties with the selection of data. Again we had a meeting with our TA, and we discussed the problem. Basically what we are going to do right now is select the long term data. So we will select products like wheat which have been measured over longer periods of time in multiple countries, in contrast to other products which have been measured over just the last couple of months and are not represented in every country. We may later in the process use this data to get more solid conclusion.

Jesper	Created some function which quickly gather information of all products, so we could quickly compare this.
Roel	Using the csv with linked products between the price dataset and the production dataset, reduced the production dataset to only include entries where the product has a link to a product in the price dataset and the country is in the price dataset. Since the country names where identical in almost all cases in both datasets, there was no need to change or link them. The production dataset is now reduced, but still needs to be cleaned.
Tessel	Working on datacleaning. Removing all nan values, and normalized units
Matthijs	Created functions for the eda, written in file called eda.py. maked sure all kinds of plots could be made, plotting price over time.

June 13, 2018 (Wednesday)

Jesper	Started with making some plots
Roel	
Tessel	Performing unit conversion for the database
Matthijs	Added additional functions to eda, and reshaped the eda.py so It could be used easier and modified better.

June 14, 2018 (Thursday)

Jesper	Searched for other dataset, find solutions which fits our problems
Roel	Checked if all units where the same in the production dataset, which they were. Removed entries in the production dataset where the production value was a NaN. Since the previous reduction was based on the country and the product, there where no NaN's in these columns. The year column had no NaN's either and other column's are not (yet) used. The production dataset is now reduced, cleaned and ready to use.
Tessel	Discussing further missing data issues with the group. Researching correlation methods and preparing data for this. Decision to discard months and turn to year averages.
Matthijs	Start researching what needs to be done for machine learning so we could make a head start on this subject. Started with some experimenting

June 15, 2018 (Friday)

Jesper	Worked further on the EDA, and start writing about it in the report
Roel	Start making scatter plots of the productiondata. Set out the price against production. But needed still to connect the two datasets with each other. Some crops have little differences in their names.
Tessel	Writing code to compute yearly averages. If one month in the data is missing it is interpolated between the previous and following month, but only if 2 or less months are missing.
Matthijs	Start with the first correlations models, still further experimenting with machinelearning

June 16, 2018 (Saturday)

Jesper	
Roel	
Tessel	
Matthijs	

June 17, 2018 (Sunday)

Jesper	finished the process for plotting the standard deviation plots. The only problem is that a 2mb html file doesn't load so well. Need to find other solutions for the bokeh plot presentation
Roel	
Tessel	
Matthijs	

June 18, 2018 (Monday)

Today we created a list of wat is done, and what need more work.

Preprocessing of price dataset is almost done, non-food products are removed, units are normalized and nan's are removed. What is left to be done is the removal of units which can't be normalized, creation of an extra data set with data per month to data with means per year so it lines up with the production dataset, and finally markets and district need normalizing to corresponding country.

Preprocessing of production dataset is reduced to the products we actually need. Null values are removed, and only the production of crops is selected from the dataset. Only a plan is needed about what to do with the missing years.

Report needs much work

Eda is in the process of making it dynamic and easy to use.

Analyse and website is currently in start-up and ideas are created.

Jesper	Implemented bokeh graphs. Adjusted the code, using bokeh instead of matplotlib. Bokeh works way better, but have some difficulties with exporting the files as png. Really difficult message to understand:

	“warnings.warn('Selenium support for PhantomJS has been deprecated, please use headless ‘”
Roel	Wrote a function that finds the product/country combinations that have the most overlapping data between both datasets (above a certain threshold), and creates production/price scatterplots for each combination using matplotlib. Since the production dataset is per year and the price dataset is per month, the yearly price needed to be computed. For now, it simply takes the mean of all available monthly prices in that year, but this will be changed in the future. Finding the best product/country combinations takes a long time, this is why the plan is to partially rewrite the function, so that all viable product/country combinations and the number of overlapping datapoints are written to a csv file, and to use this csv when plotting.
Tessel	Finishing off yearly average preprocessing. Starting on machine learning: writing code to compute product correlations
Matthijs	

June 19, 2018 (Tuesday)

Jesper	- Finding a different way of displaying graphs, solution could be found back at using flask, which helps to generate a new plot every time there is asked for one.
Roel	Made creating the price/production scatterplots a lot faster. Wrote a separate function that finds all product/country combinations with more than 1 overlapping datapoint and writes them to a csv to be used later. This function still takes a long time to run, but only needs to be run once. Using the most basic method for computing yearly price data (taking the mean of all month prices that year), this leaves about 300 combinations with 1 or more datapoints, and about 60 combinations with 10 or more datapoints. Using this csv, the scatterplots can be created fairly quickly. The plots are still made using matplotlib, the plan is to convert this to bokeh plots.
Tessel	- Working on Linear regression code
Matthijs	Not present at meeting

June 20, 2018 (Wednesday)

Jesper	- Researched for better solutions than flask
Roel	
Tessel	- Working on Linear regression code. Noted that very little significant data is available.
Matthijs	

June 21, 2018 (Thursday)

We had a meeting today. There were two main problems:

1. Trouble with machine learing, problems with the overlapping dataset. The overlapping dataset is too small to actually perform a Lineair regression Analyse.

Jesper	- Created a new repository for the website, this is needed according to github pages. The site could be found at: https://jesperss32.github.io/.
Roel	Worked on learning how bokeh works and converting the price/production scatterplots to bokeh. Created a page displaying multiple tabs with scatterplots of the product/country combinations with the most overlapping data by running a bokeh server. Since running a bokeh server can only produce pages with nothing but the plot, the plan is to find a method to include bokeh plots as a part of a html page.
Tessel	- Working on machine learning. Decision to eventually convert data to proportions.
Matthijs	Worked on the correlation models of food price which influence food prices

June 22, 2018 (Friday)

Jesper	Worked on the website, had a skype session with Roel about creating plots inside a website. Only little documentation is to be found about integrating plots into a website
Roel	Worked on the website, and had a skype session with Jesper about creating interactive plots on the website. Very difficult and not yet found what we want.
Tessel	Worked on machine learning, connecting the dots.
Matthijs	Start writing in the report about the correlation models, had some difficulties.

June 23, 2018 (Saturday)

Jesper	Created some plots for the website, figured out the layout of the website
Roel	Worked on the scatterplots for price vs production for the website.
Tessel	
Matthijs	

June 24, 2018 (Sunday)

Jesper	
Roel	Found a way to include html-snippets that are in a separate file in a html page, which is useful for things like the navigation bar, that need to be on every page. This makes the html clearer and easier to update. Searched for different methods to include a bokeh plot in a html page. The best solution (for now) is to use bokeh's autoload static function. When you run the code that produces a plot, it writes the javascript code to a separate file and prints a tag, which is a div that links to the script. Placing the tag in a html page means it will be replaced with the plot. I tried to write the tag to a file and include it in a page with the same method as the navigation bar, but for some reason this doesn't work. Instead, every you run the code for creating a plot, it prints the tag, which needs to be manually put in the page.
Tessel	
Matthijs	

June 25, 2018 (Monday)

Had a group session today after the lecture. We discussed our progress we made in the weekend and plan ahead for this week.

Jesper and Roel are working on the website, during the weekend we figured out what is the best solution for our website and how the layout should be. We discussed this at the meeting today. We dived the task up into two separate tasks. First their should be a interactive plot of food prices over time, and second should their be a plot with geographical data of food prices. Roel is working on the first task while Jesper is working on the second task.

Jesper	Merged all updates on progress-book together - Created a plot which lays out the geographical data
Roel	Worked on putting plots in the site that show the price over time for a given products and country. This now works with the product and country hard-coded, but the plan is to be able to select a country on the site and one or multiple products. Also, the time axis currently doesn't display the months and years correctly.
Tessel	Worked on correlation between productprices and production, created tables with values and run some correlation tests.
Matthijs	Worked on the correlations for productprices to productprices. Runned some test and collecting values and put this in the report.

June 26, 2018 (Tuesday)

Roel not present at meeting

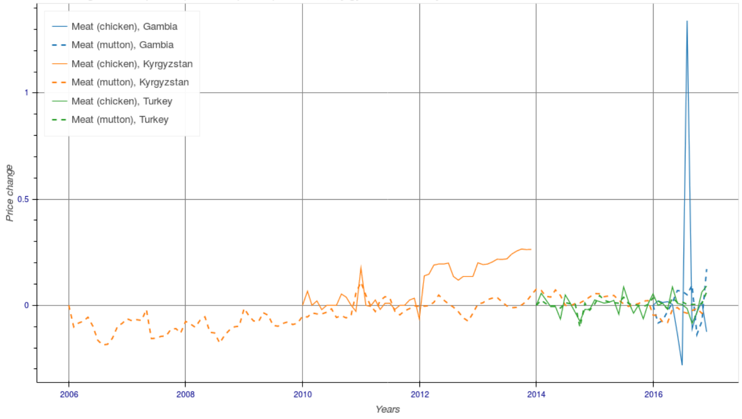
Had a meeting today were we:

- Discussed possible improvements to the report

- Asked what was the best way to create an interactive site. This appears to be using javascript callbacks.

Jesper	Worked on the geographic plots, finished the creation of them
Roel	-Worked on trying to create interaction using javascript callbacks. This should be possible, but I was not able to figure out how to work with the pandas dataframes in javascript. Instead we decided to show a selection of plots on the site that we found to be interesting.
Tessel	Worked on the report, put method and code in about the production price, and corrected parts Mathijs has written and helped Matthijs
Matthijs	Worked on the report, put the code used in the report and explained what has been done, running the final tests.

June 27, 2018 (Wednesday)

Jesper	Created mapplot hmtl page, and adjusted the plots therefore, quite some difficulties, but result was pretty ok.
Roel	Worked on creating a function that can plot the price changes over time per country for correlates products, as shown. Price changes of Meat (chicken) and Meat (mutton) in Gambia, Kyrgyzstan and Turkey  The average spearhead correlation is 0.70, taking the mean over 3 countries
Tessel	Worked on the last regression models between price and productie. Wrote the outcome date into a csv and passed it on to Jesper which is going to plot this data on the website.
Matthijs	Working on the correlation bit between products. Created therefore plot, wrote these

June 28, 2018 (Thursday)

We had a meeting today where we evaluated the results so far and made clear what things still needed to be done, mainly the results and discussion of the report, finishing the site pages we had so far and adding a home page and a page with the top 3 most interesting findings.

Jesper	Made the home page Made the page that shows the correlations between production and price change Finished the page that shows regional trends.
Roel	Finished the page of correlations between products by choosing a selection of correlations and adding text.

	Made the page with the top 3 interesting findings.
Tessel	Worked on writing method, results and discussion for the report
Matthijs	Worked on writing method, results and discussion for the report