Machine Learning Engineer Nanodegree

Capstone Project

Sunil Nayak
February 8, 2018

I. Definition

Project Overview

In this project, I tried to predict the yield of three crops (Wheat, Rice Paddy and Onion) based on factors like farm size, amount of fertilizer used, amount of water available for irrigation etc. I have also attempted to find out what factors influence the yield the most.

A United Nations study says that "The world needs to produce at least 50% more food to feed 9 billion people by 2050… Unless we change how we grow our food and manage our natural capital, food security—especially for the world's poorest—will be at risk.

http://www.un.org/en/sections/issues-depth/food/

Eradication of poverty and hunger by the year 2030 are sustainable development goals agreed by UN members. http://www.un.org/sustainabledevelopment/sustainable-development-goals/

Food security can be achieved by, increasing yield of agricultural crops, employing more land for agriculture or increasing the production of meat. It is agreed that meat is not an efficient means of food source as it takes more resources to farm animals for food consumption compared to agriculture.

Here are some studies on yield of crops:

Do small farms produce higher yield?
https://www.sciencedirect.com/science/article/pii/0305750X85900543

Yield trends are insufficient to feed the world by 2050:
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0066428

This project was inspired by this research study:
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0156571

Problem Statement

The intent of this project is to predict the yield of crops based on inputs like farm size, amount of fertilizer used, amount of water used, amount of labor and farm equipment used etc. The yield of crop varies considerably from region to region even within the same country. In this project I use data that is available at country level. I did not use more granular data for various regions within countries.

Food and Agriculture Organization of United Nations provides agricultural data by country:

http://www.fao.org/faostat/en/#data

From this website I use the data for crop yield, water used, fertilizer used, pesticide used, soil erosion and degradation. Most of the data is provided by country, year and per hectare of land.

Nation Master is another website that provides data sourced from CIA World Factbook (https://www.cia.gov/library/publications/the-world-factbook/)

http://www.nationmaster.com/country-info/groups/High-income-OECD-countries/Agriculture

From here I use the data for region and income category a country belongs to, farm worker count, tractor count. This data is provided by country, year and per hectare of land.

The tasks involved were:

- Download agricultural data from various sources. Process this data: eliminate factors for which we have insufficient data. Estimate missing data where appropriate.
- Train couple of regression models.
- Find the factors that influence the yield the most.

Metrics

The performance of the model is by these methods:

1. *RMSE: Root mean squared error of yield*
   RMSE = square root [sum((predicted yield– observed yield)^2) / n]
   Where n is the number of observations
   We can see RMSE is what fraction of average yield.
2. Coefficient of determination (R squared value): the ratio of explained variation to total variation of yield:
   r-squared = 1 – [SS_residual / SS_total)
   SS_residual: sum[(predicted yield – observed yield)^2]
   This is the measure of error when comparing predicted yield from model with the observed yield.
   SS_total: sum[(observed yield – mean(observed yield))^2]
   This is the measure of error when comparing the observed yield with the model that always predicts the average yield.
   This value ranges between 0 and 1 and higher value indicates a getter fit
3. Graphs of predicted yield Vs actual yield.

II. Analysis

Data Exploration

Data fields explained

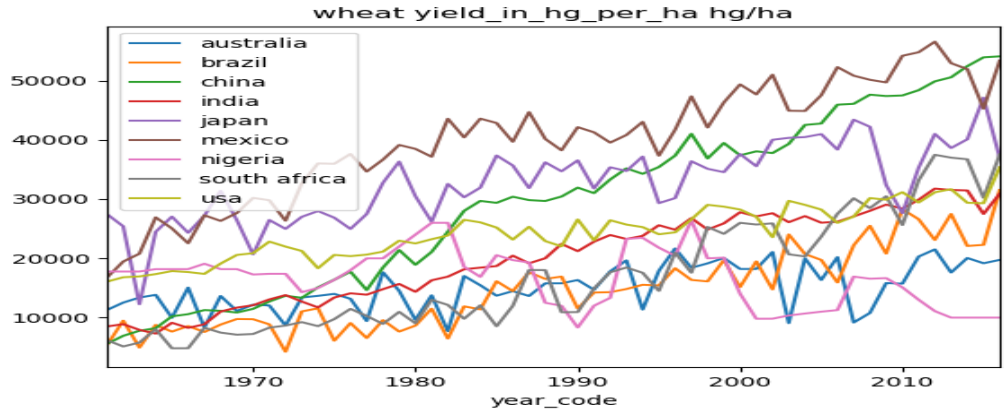
Field	Name	Short Name	Source	Type	Unit
Yield	yield	yield	FAO	Float	Hectogram/hectare
Year	year_code	year_code	FAO	Integer	yyyy
Country Code	area_code	area_code	FAO	Integer	
Crop code	item_code	item_code	FAO	Integer	

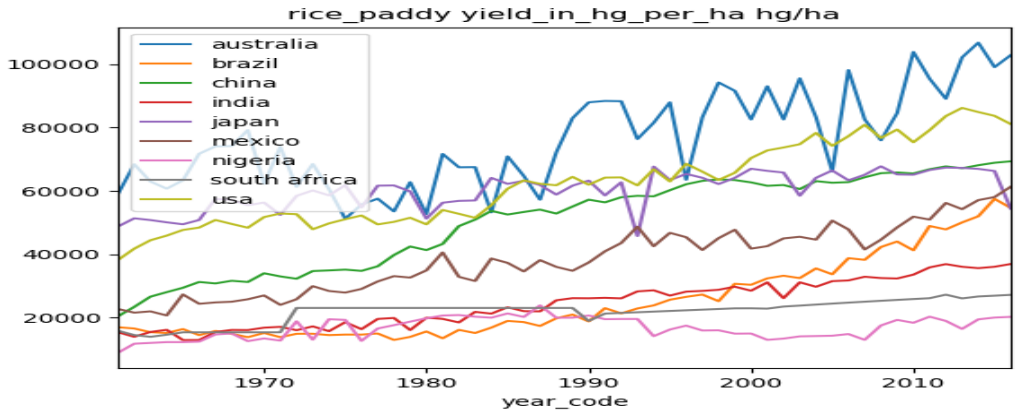
Region	region_code	region_code	NM	Integer	
Income	income_code	income_code	NM	Integer	
Fertilizer Nitrogen	fert_used_kg_per_ha_Nitrogen	F_N	FAO	Float	Kg/hectare
Fertilizer Phosphate	fert_used_kg_per_ha_Phosphate	F_P	FAO	Float	Kg/hectare
Fertilizer Potash	fert_used_kg_per_ha_Potash	F_K	FAO	Float	Kg/hectare
Arable land	percent_agg_land_arable_land	L_AL	FAO	Float	% of total agricultural land
Permanent	percent_agg_land_permanent_crops	L_PC	FAO	Float	% of total agricultural land
Permanent meadows	percent_agg_land_permanent_meadows	L_PM	FAO	Float	% of total agricultural land
Irrigated land	percent_agg_land_irrigated	L_I	FAO	Float	% of total agricultural land
Land equipped for irrigation	percent_agg_land_equiped_for_irrigation	L_EI	FAO	Float	% of total agricultural land
Pesticide used	pesticide_average_use_in_kg_per_ha	Pest	FAO	Float	Kg/ha
Water used in agriculture	water_used_in_agri_as_perc_of_total_water_used	W	FAO	Float	% of total water used
GDP per capita	gdp_per_capita_usd	GDP	NM	Float	USD
Energy used in agriculture	perc_of_total_energy_used_for_agri	Enrg	FAO	Float	% total energy used
Soil Erosion	soil_erosion_average_in_GLASOD_degrees	S_Ero	FAO	Float	GLASOD degrees *
Soil degradation	soil_degradation_average_in_GLASOD_degrees	S_Deg	FAO	Float	GLASOD degrees*
Carbon in top soil	carbon_in_topsoil_average_perc_in_weight	C	FAO	Float	% in weight
Country Name	country_name	country_name	FAO	Text	
Farms < 1 ha in size	lt_1_ha	lt_1h	NM	Float	% of total number of farms
Farms between 1 and 2 ha	1_2_ha	1_2h	NM	Float	% of total number of farms
Farms between 2 and 5 ha	2_5_ha	2_5h	NM	Float	% of total number of farms
Farms between 5 and 10 ha	5_10_ha	5_10h	NM	Float	% of total number of farms
Farms between 10 and 20 ha	10_20_ha	10_20h	NM	Float	% of total number of farms
Farms between 20 and 50 ha	20_50_ha	20_50h	NM	Float	% of total number of farms
Farms between 50 and 100 ha	50_100_ha	50_1hh	NM	Float	% of total number of farms
Farms between 100 and 200 ha	100_200_ha	1h_2hh	NM	Float	% of total number of farms
Farms between 200 and 500 ha	200_500_ha	2h_5hh	NM	Float	% of total number of farms
Farms between 500 and 1000 ha	500_1000_ha	5h_1kh	NM	Float	% of total number of farms
Farms > 1000 ha	gt_1000_ha	gt_1kh	NM	Float	% of total number of farms
Workers	workers_per_ha	Wkr	NM	Float	count / ha
Tractors	tractors_per_100_ha	Trk	NM	Float	count / 100 ha

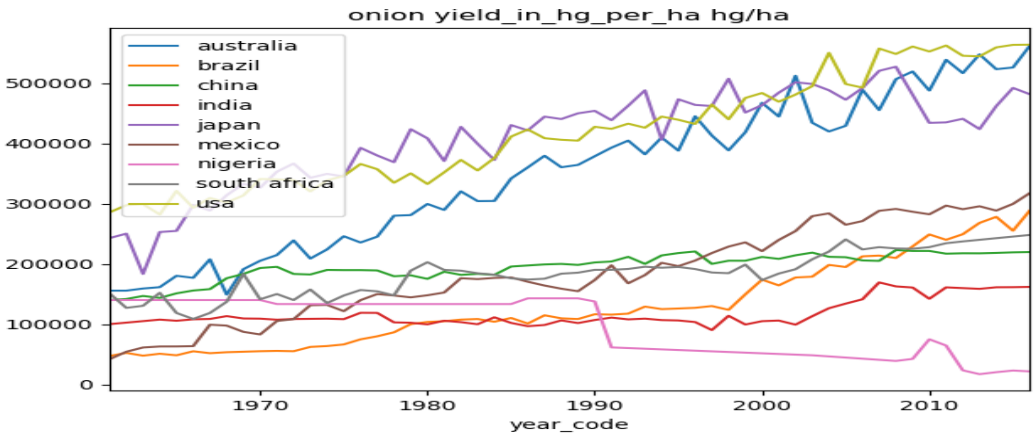
Plotting the data for each field

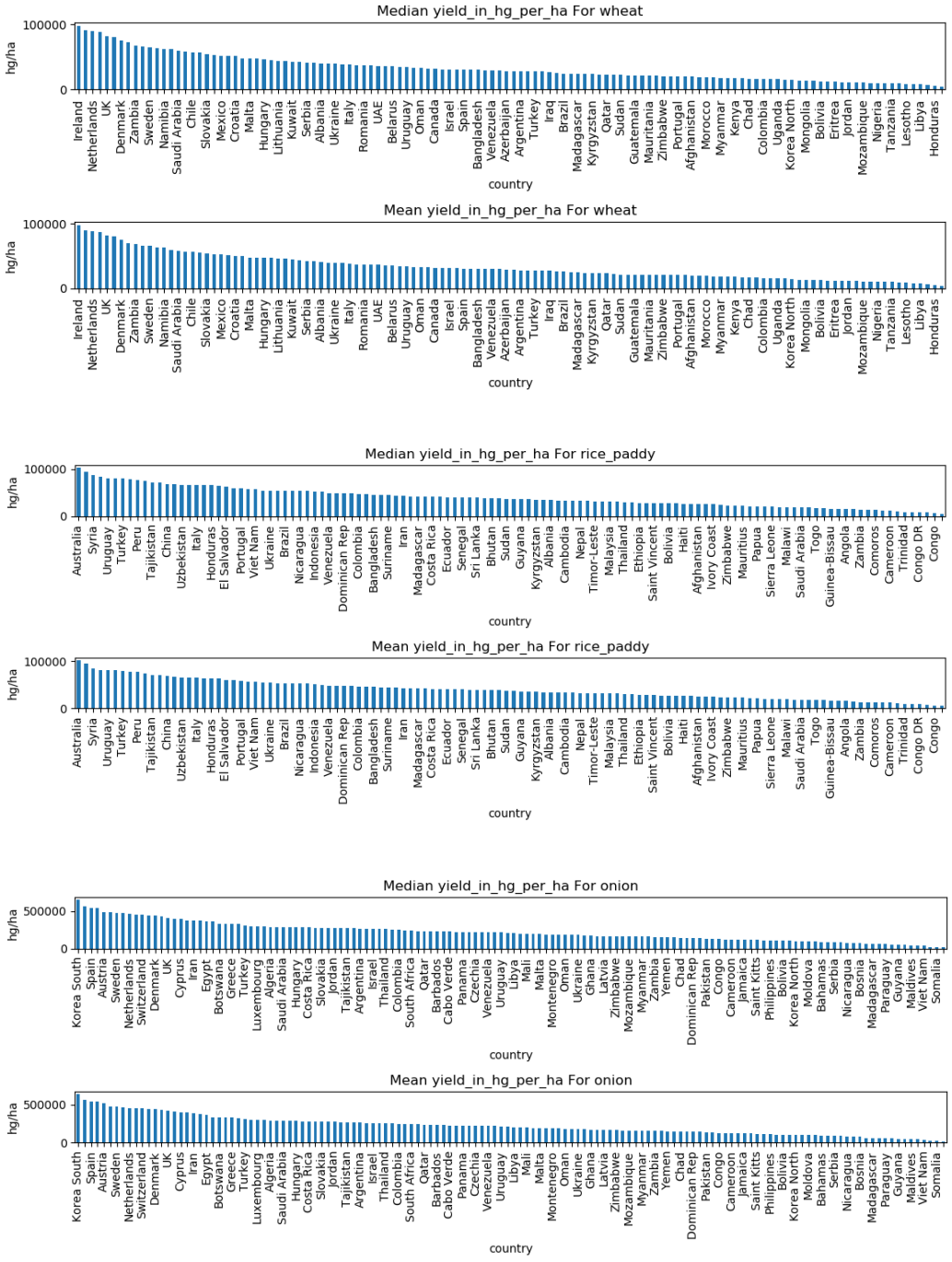
Yield (hectogram per hectare):

Here are the yields for wheat, rice paddy and onion for select countries. No one country leads in the yield for all the three crops. We have yield for all countries for the last 40 years or more.



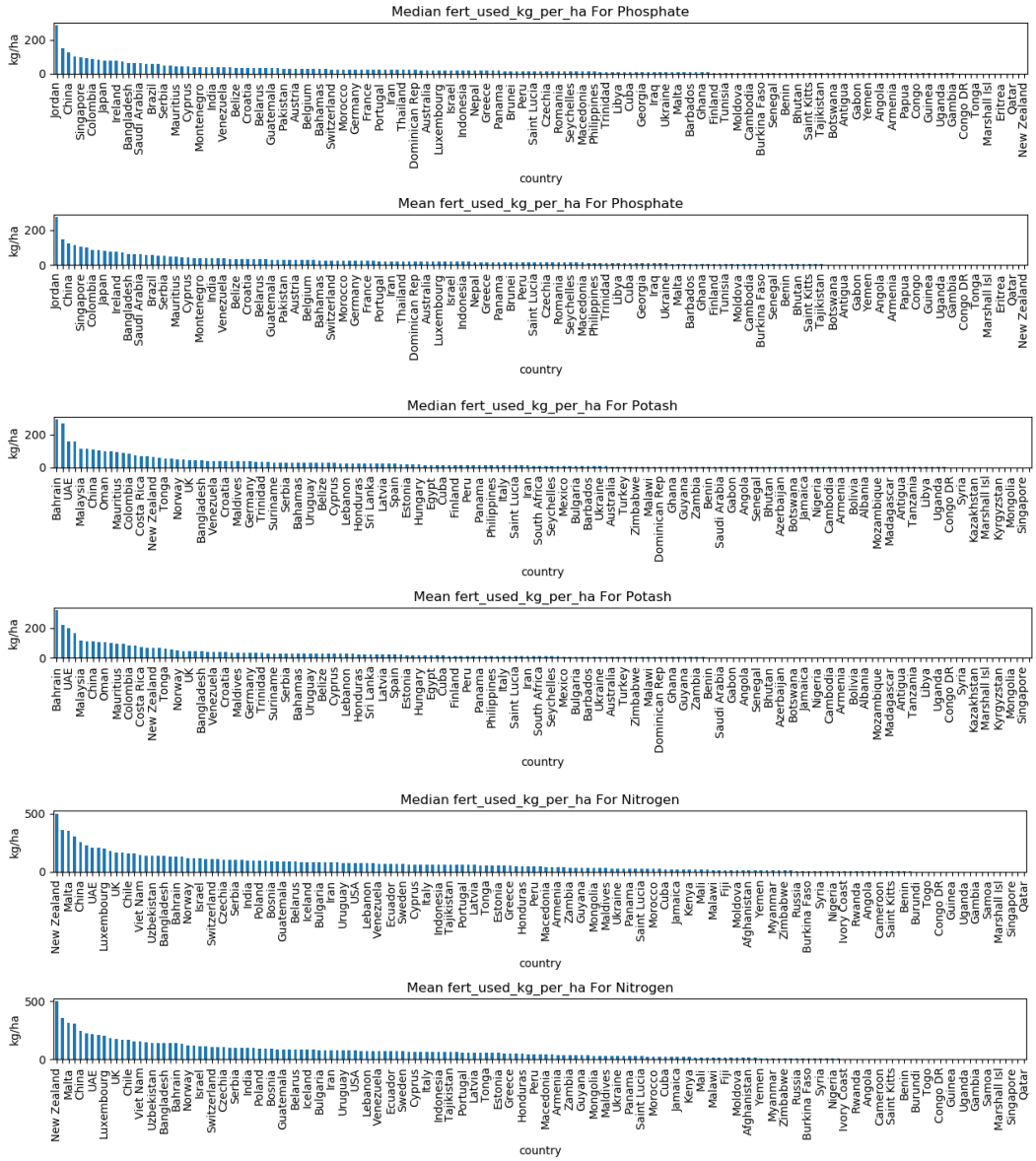






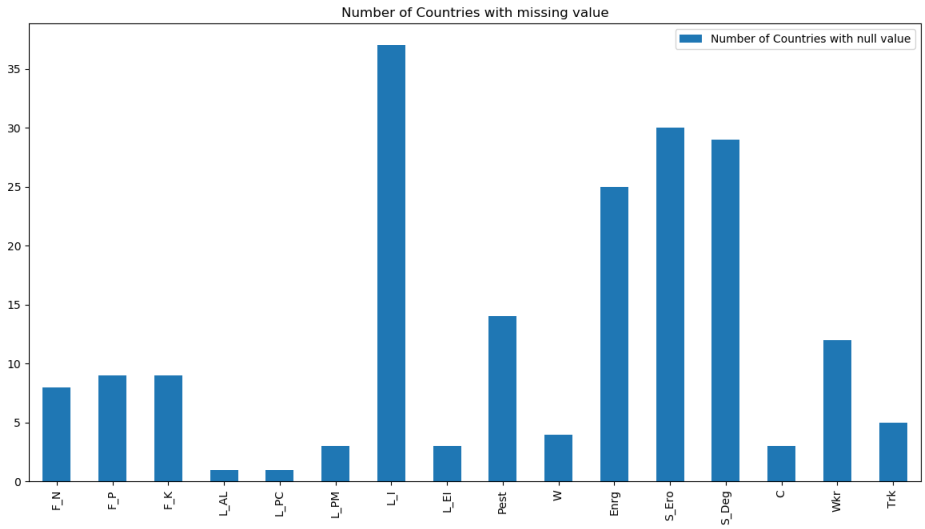
Different countries lead in the yield for each of the three crops. A single model may not capture the inter crop variation. I create one model for each crop.

Fertilizer used:

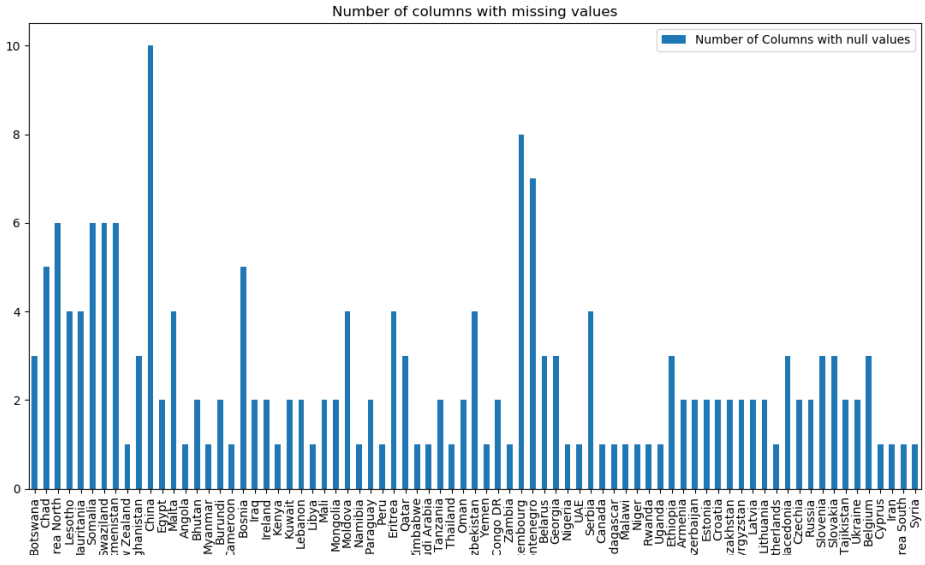


Missing data analysis

Looking at the data it appears that we do not have data for some fields:



We do not have enough data for irrigated land (L_I), energy used (Enrg), soil erosion (S_Ero) and soil degradation (S_Deg).



China, Luxembourg, Montenegro etc. are not reporting data for many of the fields. So we will end up eliminating those countries from our study.

Logic for eliminating fields and countries with insufficient data:

- Eliminate the field for which we do not have data for highest number of countries.

- Eliminate the country for which we do not have data for highest number of fields.
- Continue above steps until we only have fields that are missing data for less than 15 countries and we have countries that are missing data for fewer than 5 fields.

Filling missing data

The data is sorted by country, item and year. I fill missing data using code:

```
# based on output of this we know we have data between 2002 and 2015
get_max_min_year_with_values('F_N')
# get data only for years 2002 - 2015
all_df = all_df[all_df['year_code'].isin(range(2002, 2016))]
all_df = all_df[~all_df['area_code'].isin(eliminated_countries)]
all_df = all_df.drop(eliminated_columns, axis=1)
# fill na for columns (only inner missing values)
columns = set(data_columns) - set(eliminated_columns)    fill_na_within_time_series(columns)
# ['F_N', 'F_P', 'F_K', 'L_AL', 'L_PC', 'L_PM', 'L_I', 'L_EI']
# land use does not change dramatically year over year
# it is OK to ffill and bfill these values
fill_na_at_ends_of_time_series(columns)
#['L_AL', 'L_PC', 'L_PM', 'L_I', 'L_EI']
```

Drop rows with missing data

We have now removed countries and fields with insufficient data. We have also filled data for fields with values from nearest row. Now we can eliminate rows with missing data.

```
def drop_na():
    all_df.dropna(inplace=True)
```

Fields after data cleanup

['item_code', 'area_code', 'year_code', 'yield', 'region_code', 'income_code', 'F_N', 'F_P', 'F_K', 'L_AL', 'L_PC', 'L_PM', 'L_EI', 'W', 'GDP', 'C', 'country_name', 'lt_1h', '1_2h', '2_5h', '5_10h', '10_20h', '20_50h', '50_1hh', '1h_2hh', '2h_5hh', '5h_1kh', 'gt_1kh', 'Wkr', 'Trk']

item_code, area_code are key columns
region_code, income_code are categorical columns that are not input fields to the regression model. They are used for some comparison reporting.
year_code, country_name and GDP is not an input field to the regression model either.

Countries that remain after data cleanup

Country	Region	Income
Albania	Europe and Central Asia	Upper-middle-income
Austria	High-income	High-income
Brazil	Latin America and the Caribbean	Upper-middle-income

Bulgaria	Europe and Central Asia	Upper-middle-income
Chile	Latin America and the Caribbean	Upper-middle-income
Congo DR	Sub-Saharan Africa	Low-income
Croatia	High-income	High-income
Czech Republic	High-income	High-income
Ecuador	Latin America and the Caribbean	Upper-middle-income
Estonia	High-income	High-income
Ethiopia	Sub-Saharan Africa	Low-income
Finland	High-income	High-income
France	High-income	High-income
Georgia	Europe and Central Asia	Lower-middle-income
Germany	High-income	High-income
Greece	High-income	High-income
Guatemala	Latin America and the Caribbean	Lower-middle-income
Honduras	Latin America and the Caribbean	Lower-middle-income
Hungary	High-income	High-income
India	South Asia	Lower-middle-income
Ivory Coast	Sub-Saharan Africa	Lower-middle-income
Jordan	Middle East and North Africa	Upper-middle-income
Kyrgyzstan	Europe and Central Asia	Low-income
Latvia	Europe and Central Asia	Upper-middle-income
Lebanon	Middle East and North Africa	Upper-middle-income
Lithuania	Europe and Central Asia	Upper-middle-income
Malawi	Sub-Saharan Africa	Low-income
Namibia	Sub-Saharan Africa	Upper-middle-income
Nepal	South Asia	Low-income
Netherlands	High-income	High-income
Norway	High-income	High-income
Pakistan	South Asia	Lower-middle-income
Philippines	East Asia and the Pacific	Lower-middle-income
Poland	High-income	High-income
Portugal	High-income	High-income
Romania	Europe and Central Asia	Upper-middle-income
Senegal	Sub-Saharan Africa	Lower-middle-income
Spain	High-income	High-income
Sweden	High-income	High-income
Thailand	East Asia and the Pacific	Upper-middle-income
Togo	Sub-Saharan Africa	Low-income
Turkey	Europe and Central Asia	Upper-middle-income
Uganda	Sub-Saharan Africa	Low-income
UK	High-income	High-income

USA	High-income	High-income
Vietnam	East Asia and the Pacific	Lower-middle-income
Yemen	Middle East and North Africa	Lower-middle-income
Zambia	Sub-Saharan Africa	Lower-middle-income

Country count by region

Region	Country Count
East Asia and the Pacific	3
Europe and Central Asia	8
High-income	17
Latin America and the Caribbean	5
Middle East and North Africa	3
South Asia	3
Sub-Saharan Africa	9
Grand Total	**48**

Country count by income

Income	Country count
High-income	17
Low-income	7
Lower-middle-income	11
Upper-middle-income	13
Grand Total	**48**

We end up selecting 17 high-income countries, which makes more than one third of all the countries selected. The reason for this could be because the high-income countries monitor and report their data better. Attributes that influence the yield in high-income countries may dominate the model.

Variability In Data

There is variability in some of the fields like Fertilizer used and number of tractors used. That is expected as some countries (probably with lax laws) use a lot of potentially toxic fertilizers and some countries (richer with less population) tend to use lot of tractors and have a highly automated farming industry. GDP is listed just to show the disparity in the wealth among countries.

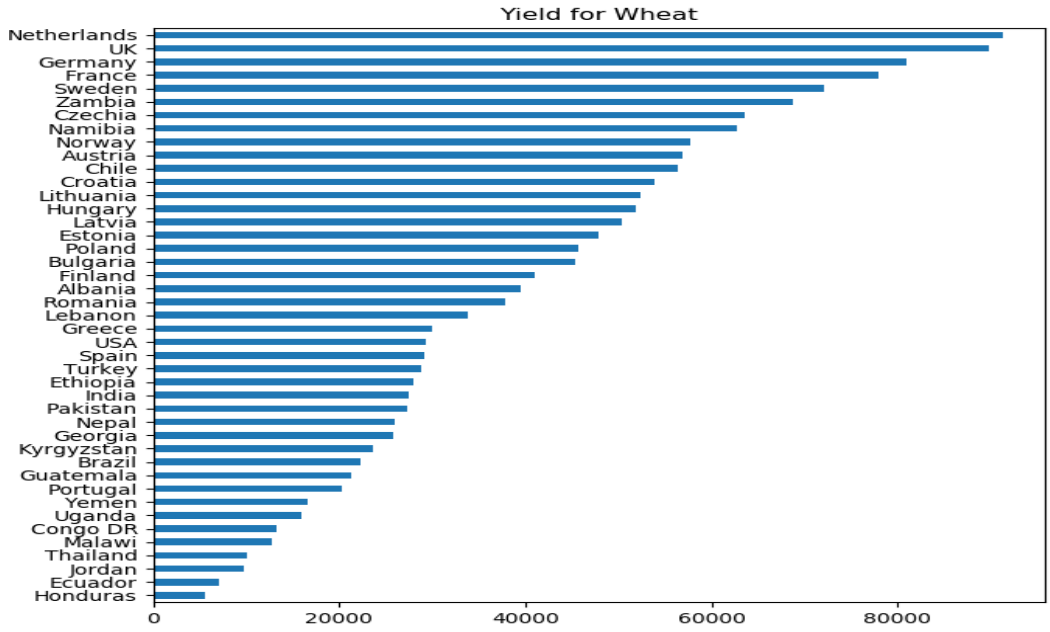
	F_N	F_P	F_K	GDP	Trk
count	1610.00	1610.00	1610.00	1610.00	1610
mean	62.93	24.29	21.76	13042.05	381.685043
std	50.97	35.39	38.42	16745.91	466.179124
min	0.00	0.00	0.00	109.26	0.32
25%	25.90	6.70	1.84	1215.30	38.65
50%	56.40	17.07	12.39	5349.68	180.42

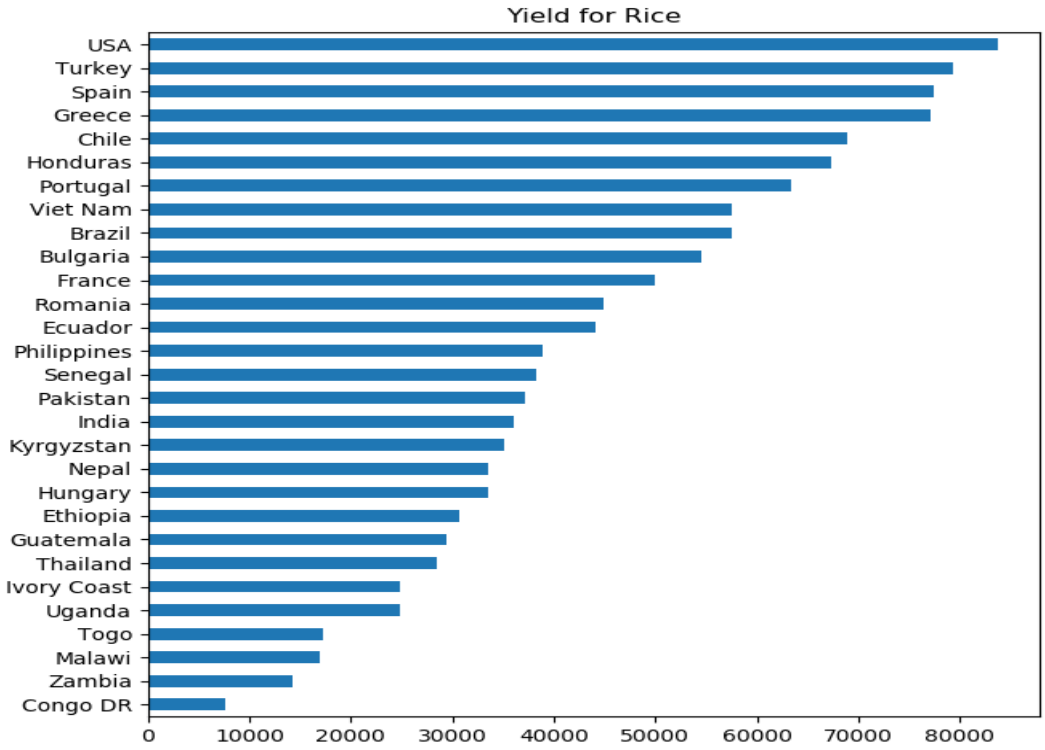
75%	81.46	29.01	28.18	17561.71	670.82
max	313.85	405.55	414.67	102832.96	2379.24

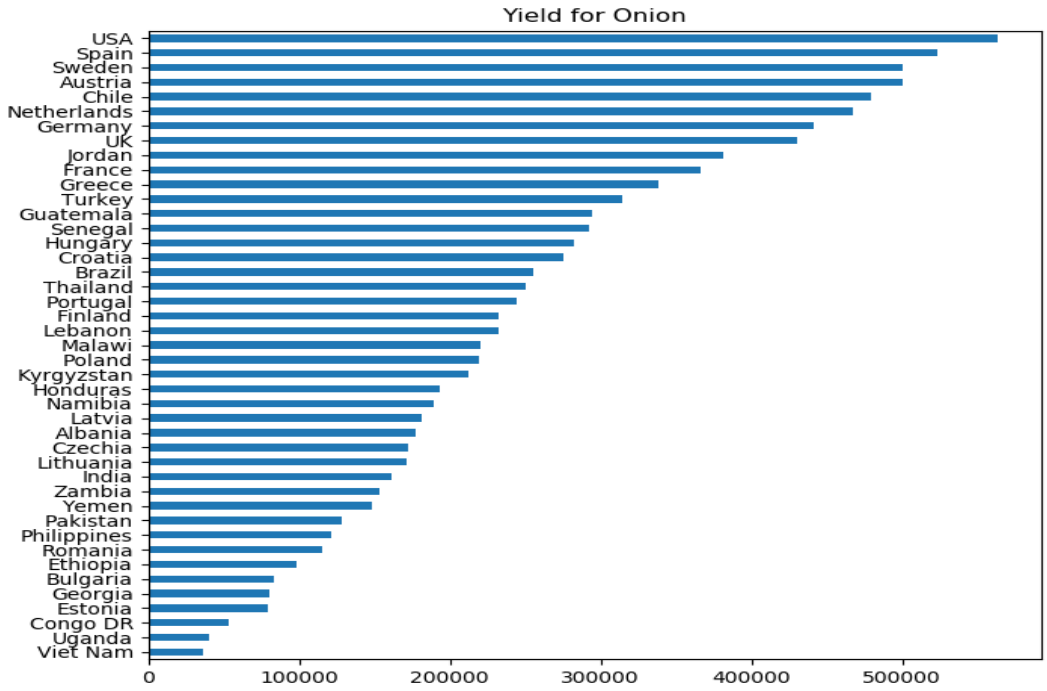
Exploratory Visualization

Here is some exploratory visualization on sanitized data.

Yield in hectogram (100 grams) per hectare looks for different countries in year 2015:

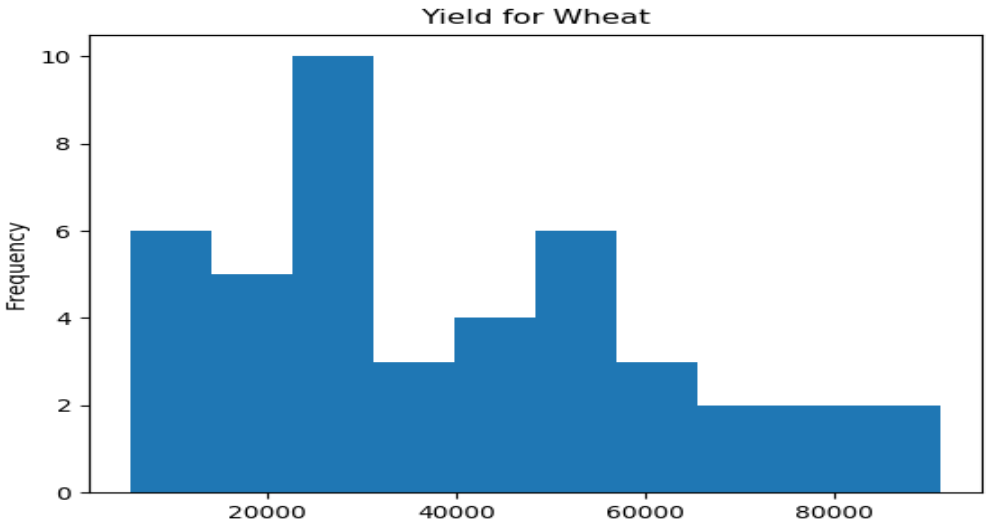



Yield for Rice
USA
Turkey
Spain
Greece
Chile
Honduras
Portugal
Viet Nam
Brazil
Bulgaria
France
Romania
Ecuador
Philippines
Senegal
Pakistan
India
Kyrgyzstan
Nepal
Hungary
Ethiopia
Guatemala
Thailand
Ivory Coast
Uganda
Togo
Malawi
Zambia
Congo DR
0
10000
20000
30000
40000
50000
60000
70000
80000

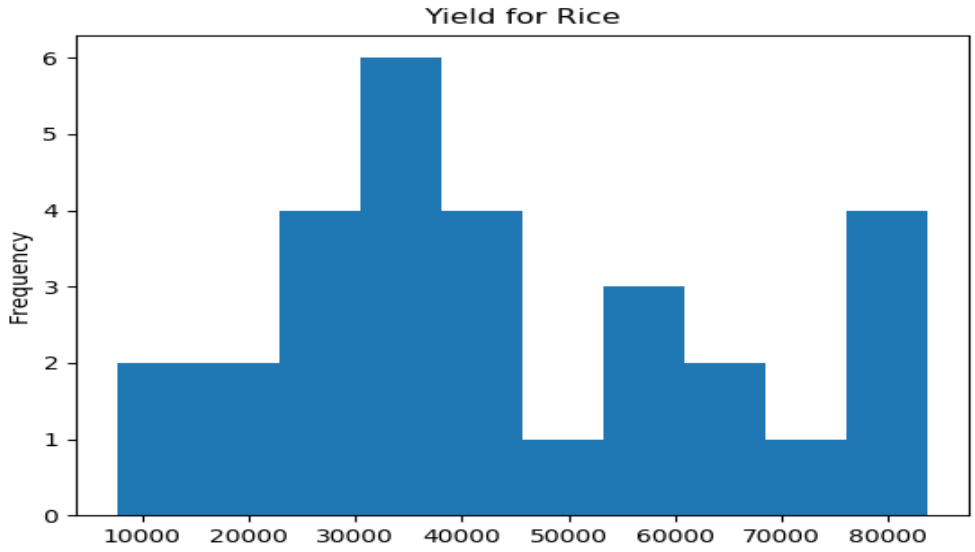


Not all countries grow all crops. The same country does not lead in yield of all crops.

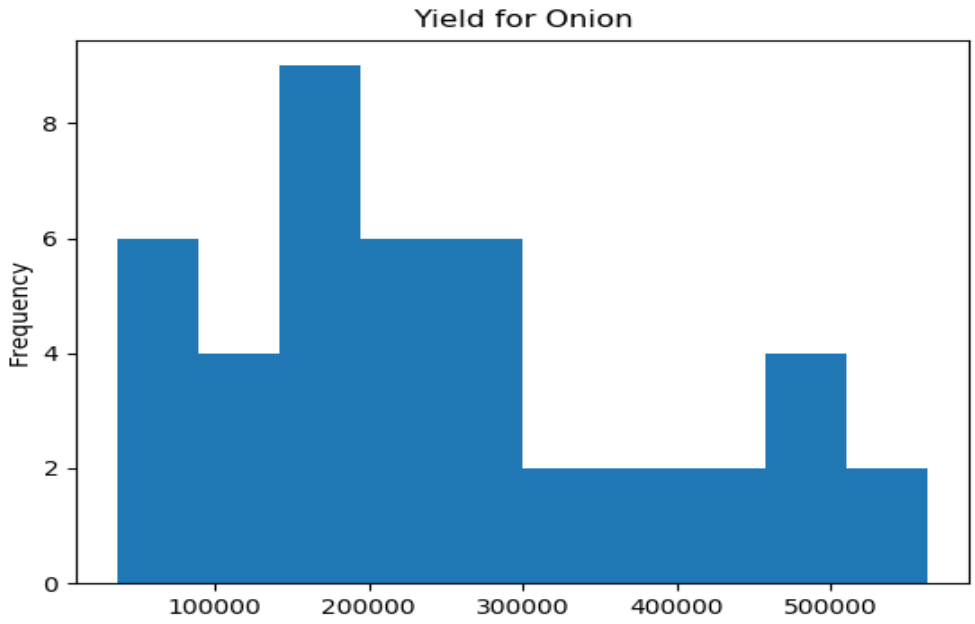
Yield distribution:



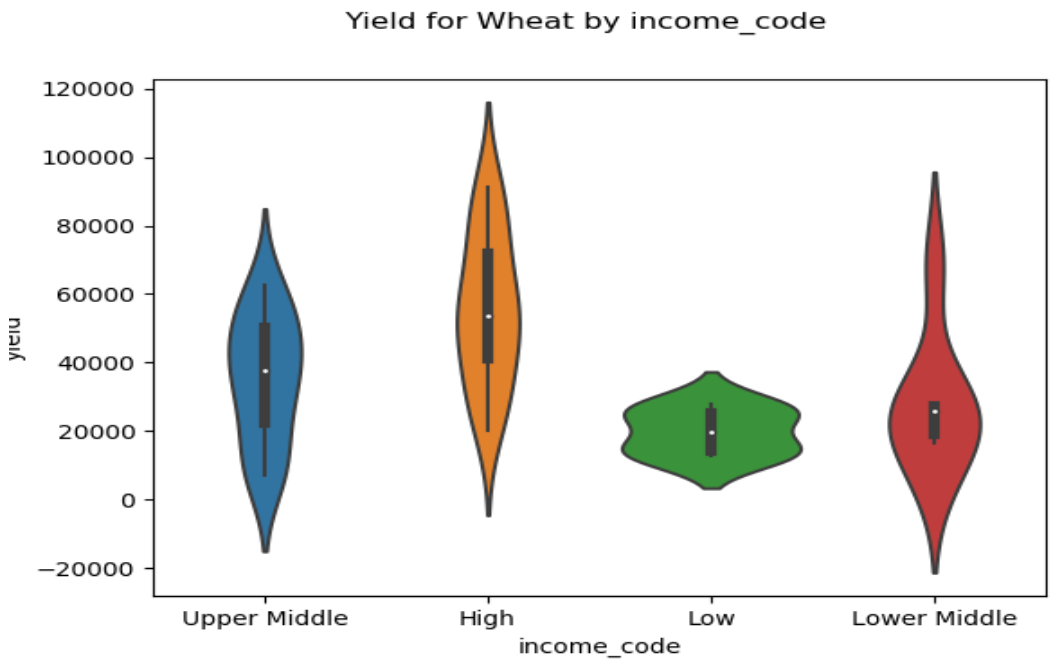
Most countries have lower yield of wheat. Very few countries have a yield of 50,000 hectogram or above per hectare.



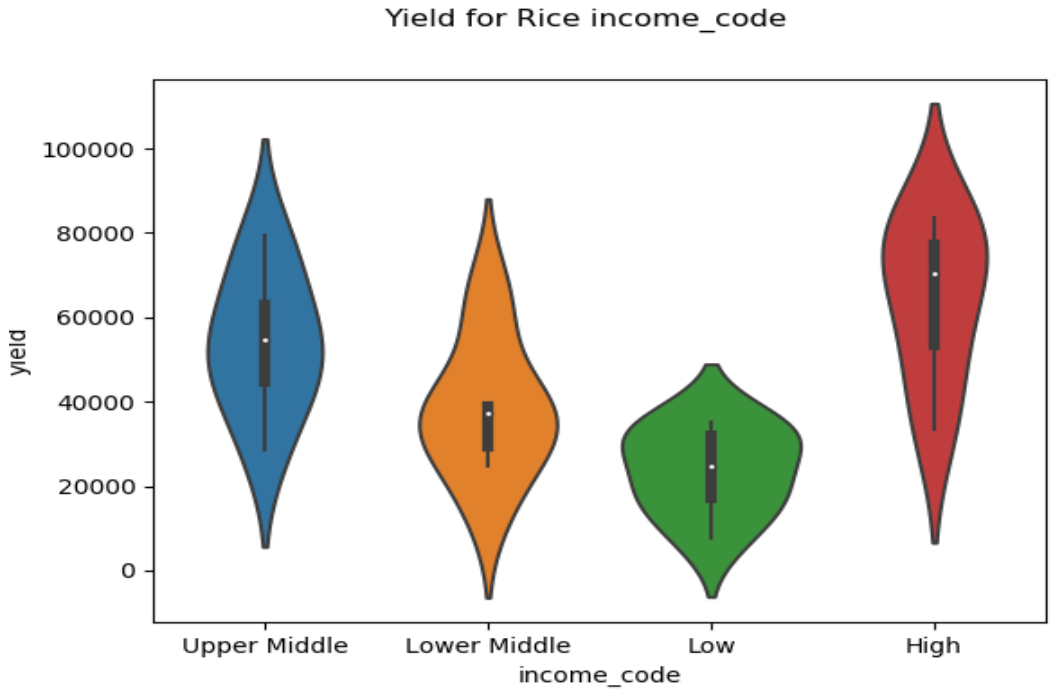
Most countries produce close to median yield for rice. Few countries have much higher yield than others (80,000 hectogram/hectare).



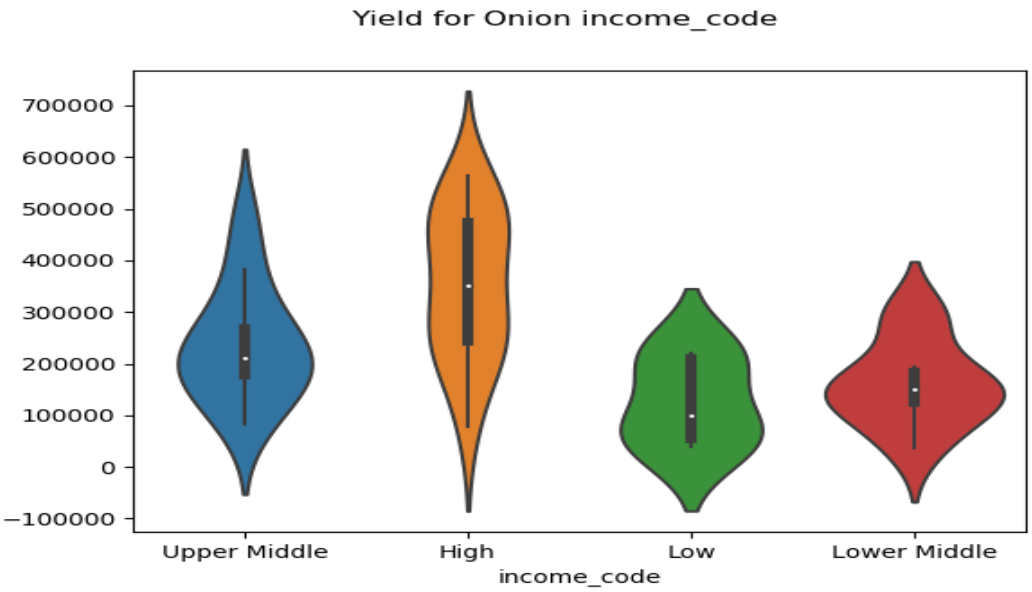
Most countries have a yield for onion on the lower end. Very few have very high yield (500,000 hectograms per hectare).



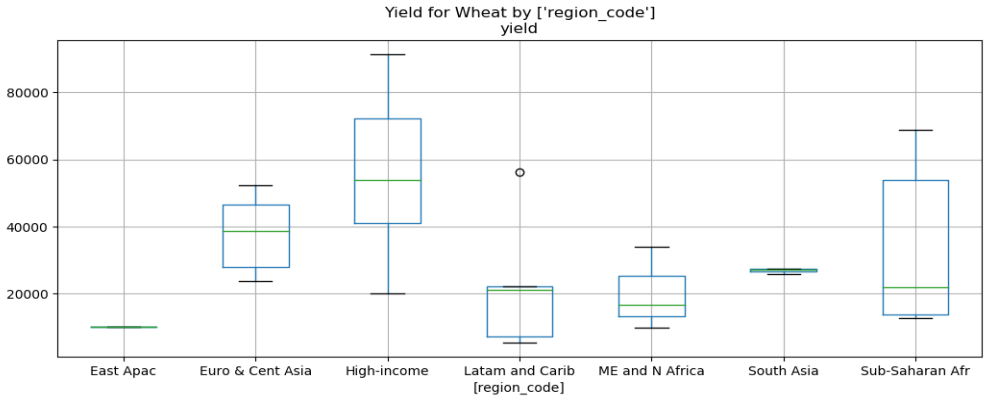
There appears to be some correlation between the income of the country and the yield of wheat. Higher income countries have better yield on average.



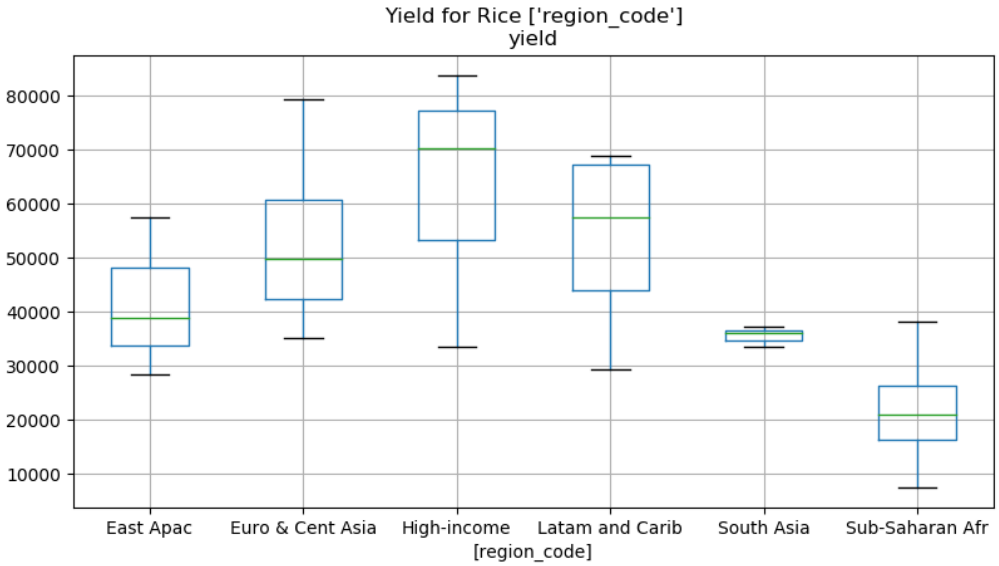
The same trend continues for yield of rice. Higher income countries have higher yield than other countries.



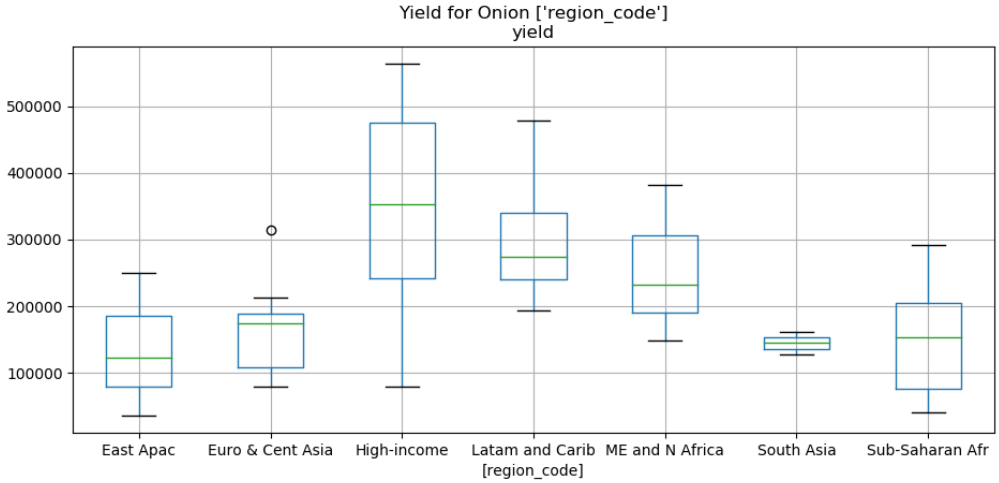
Same for yield of onion.



High-income countries are considered in a separate region. Europe produces higher yield of wheat than other geographic regions. In Latin American And Caribbean region Brazil (outlier) has a much higher yield of wheat than other countries.

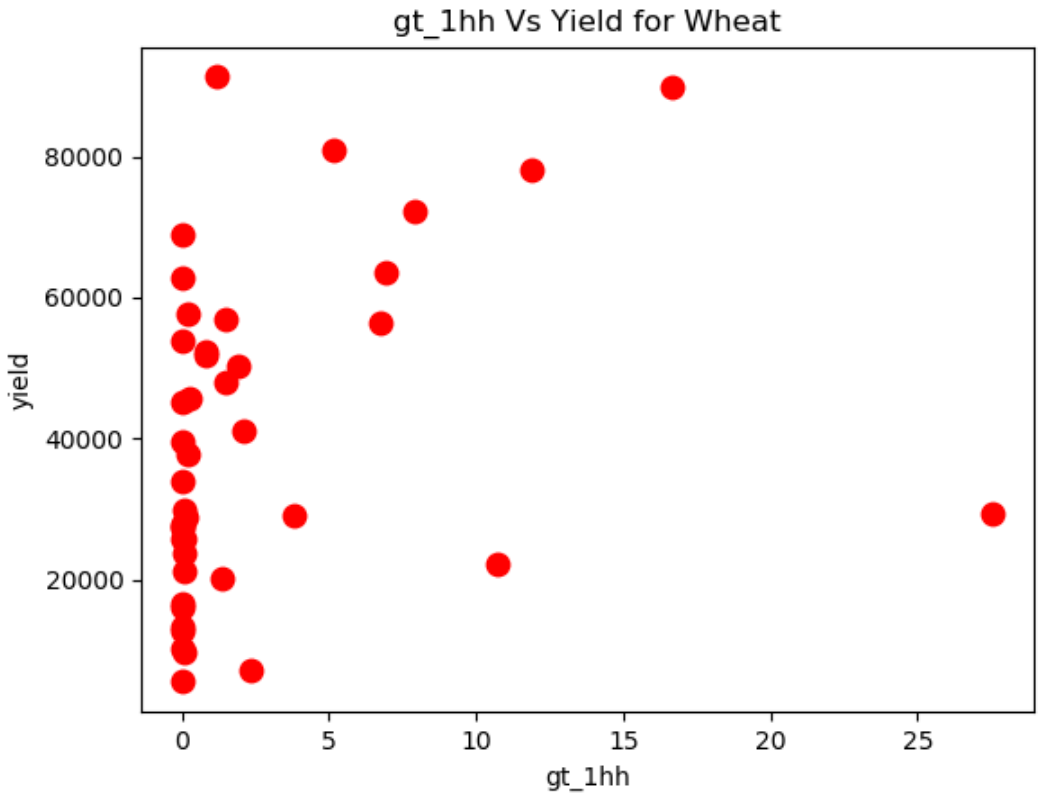


High-income countries and Latin American countries lead the pack.



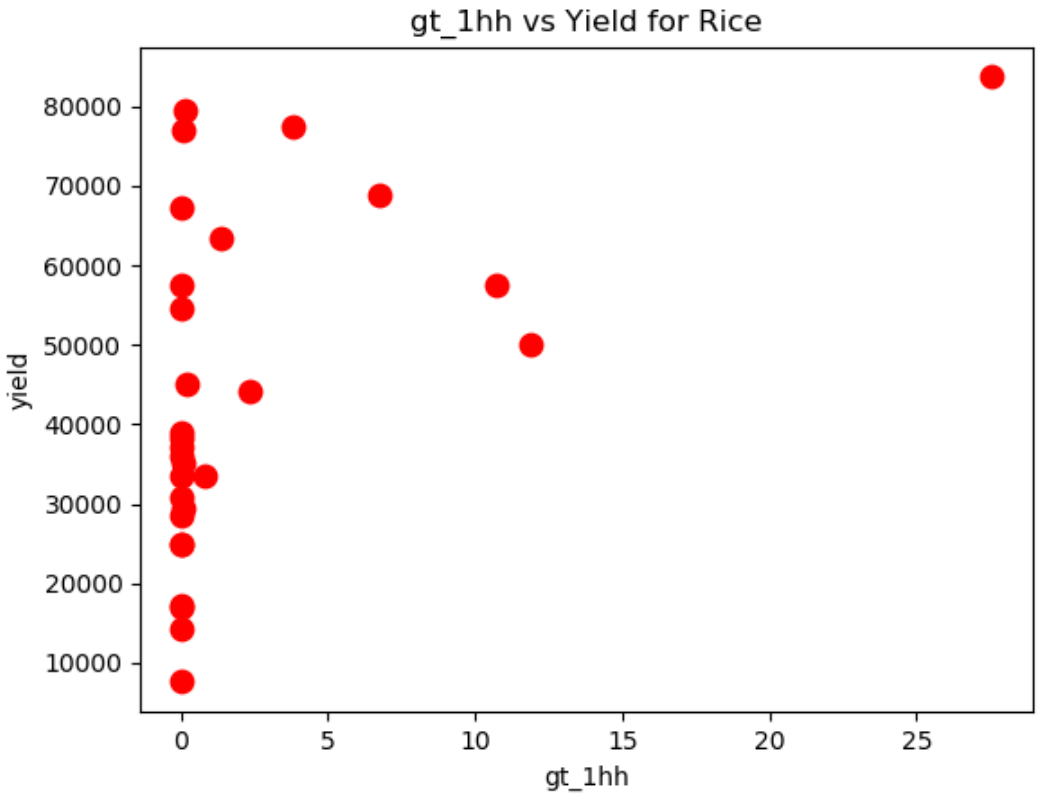
In Europe & Central Asia, Turkey (outlier) leads the yield for Onion by a huge margin.

Yield by various farm sizes:



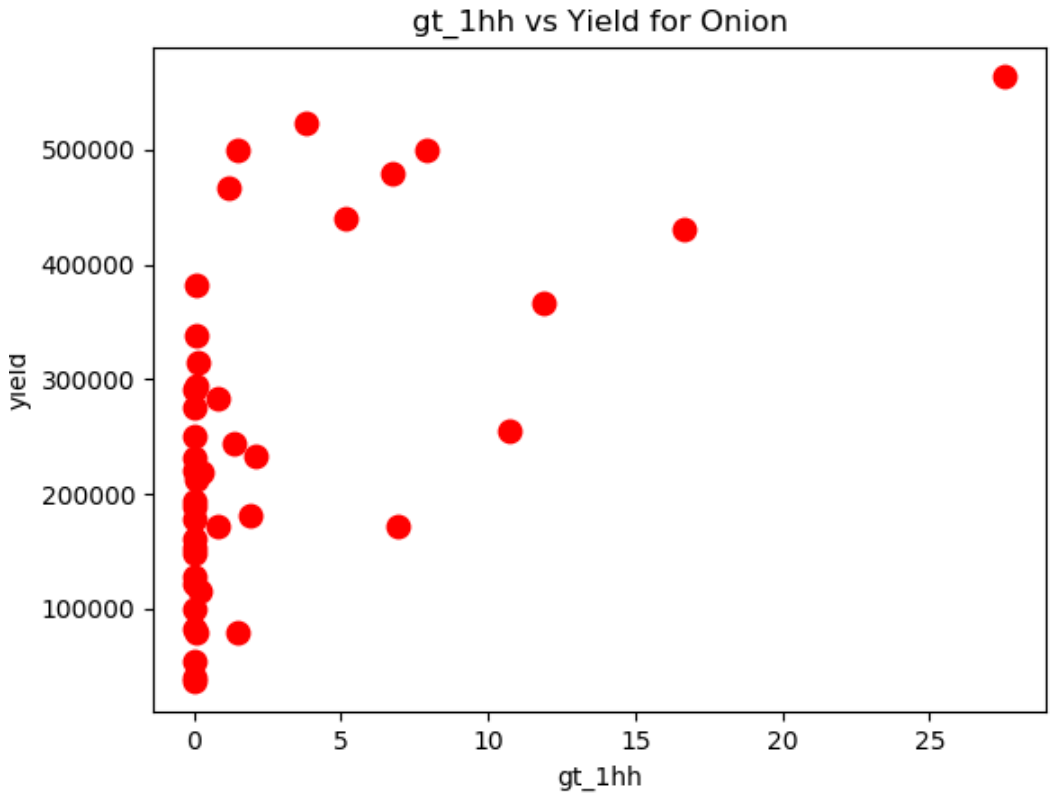
X-axis is % of farms of size greater than 100 hectares. Most countries have no farm greater than 100 hectares.

Countries having farm size of more than 100 hectares have high yield of wheat but for two outliers. I have combined all columns for field size greater than 100 hectares into one field because the data for field size 100-200, 200-500, 500-100, > 100 hectares is 0 for many cases.



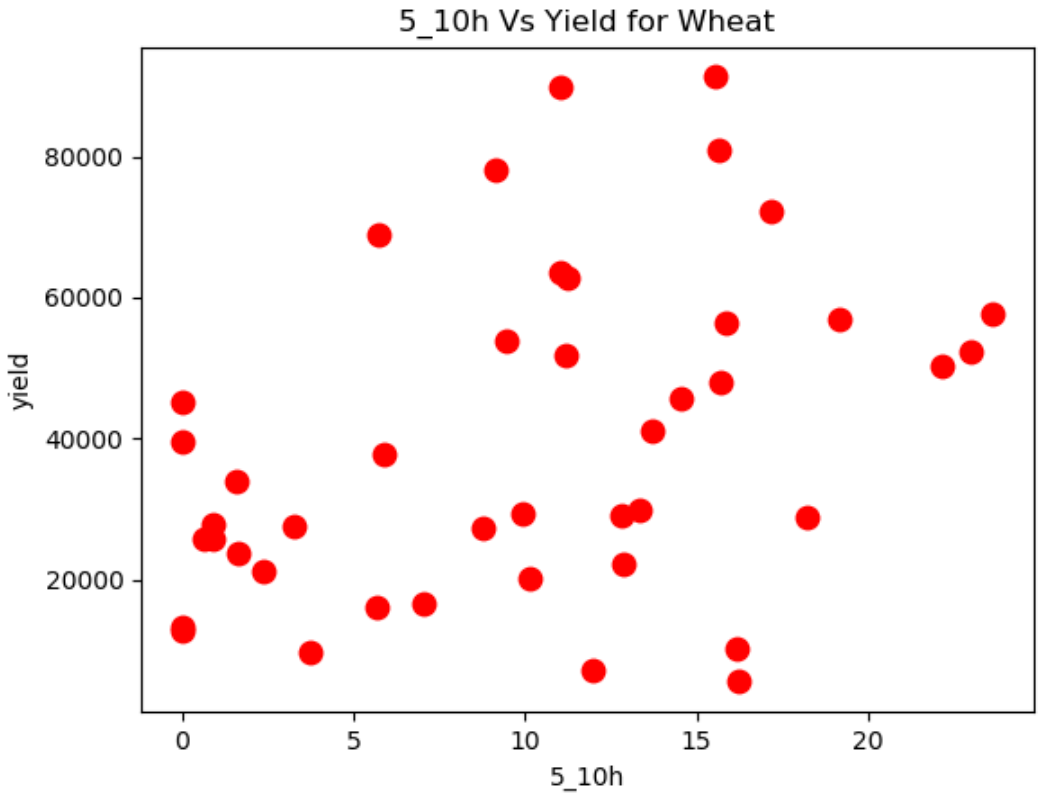
X-axis is % of farms of size greater than 100 hectares. Most countries have no farm greater than 100 hectares.

Countries having farm size greater than 100 hectares are not necessarily have very high yield of rice.



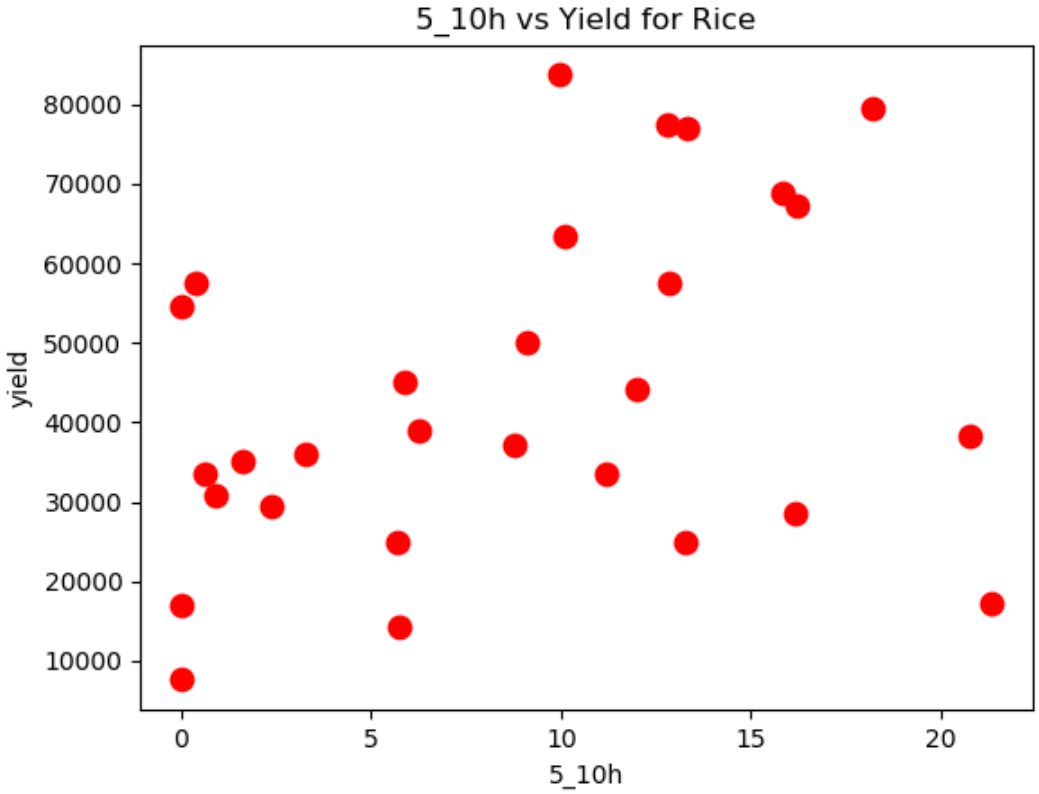
X-axis is % of farms of size greater than 100 hectares. Most countries have no farm greater than 100 hectares.

Countries having farm sizes greater than 100 hectares not necessarily have high yield of Onion.



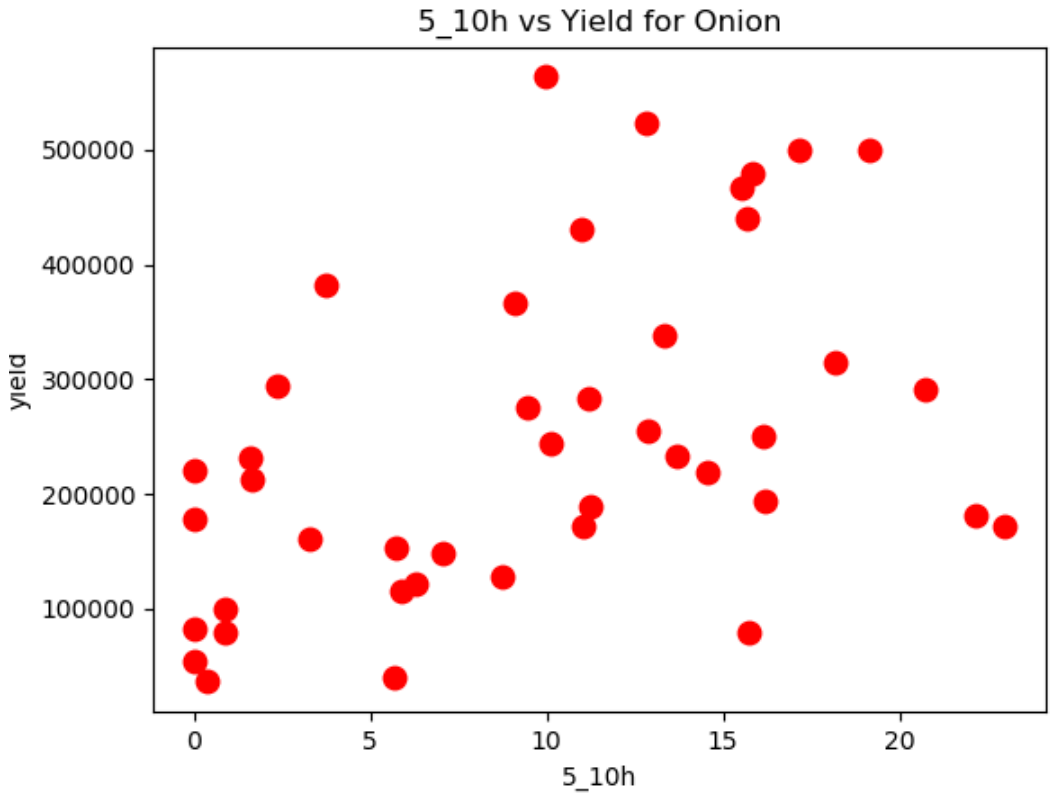
X-axis is % of farms with size between 5 and10 hectares. Most countries have no farm greater than 100 hectares.

No trend in this graph. Countries that have 10 -15% of farms between 5 and 10 hectares also show low yield like countries that have fewer than 5% of farms between 5 and 10 hectares.



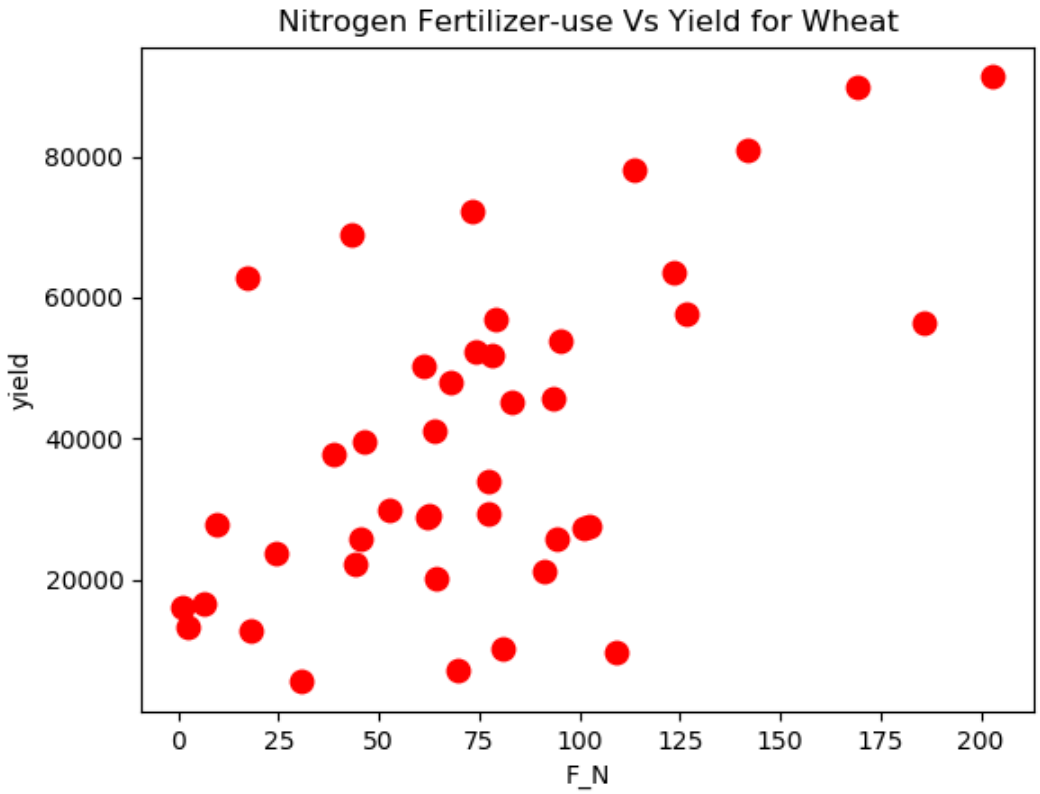
X-axis is % of farms with size between 5 and10 hectares. Most countries have no farm greater than 100 hectares.

Countries with high % of farms of size 5-10 hectares also have high yield for rice in general.



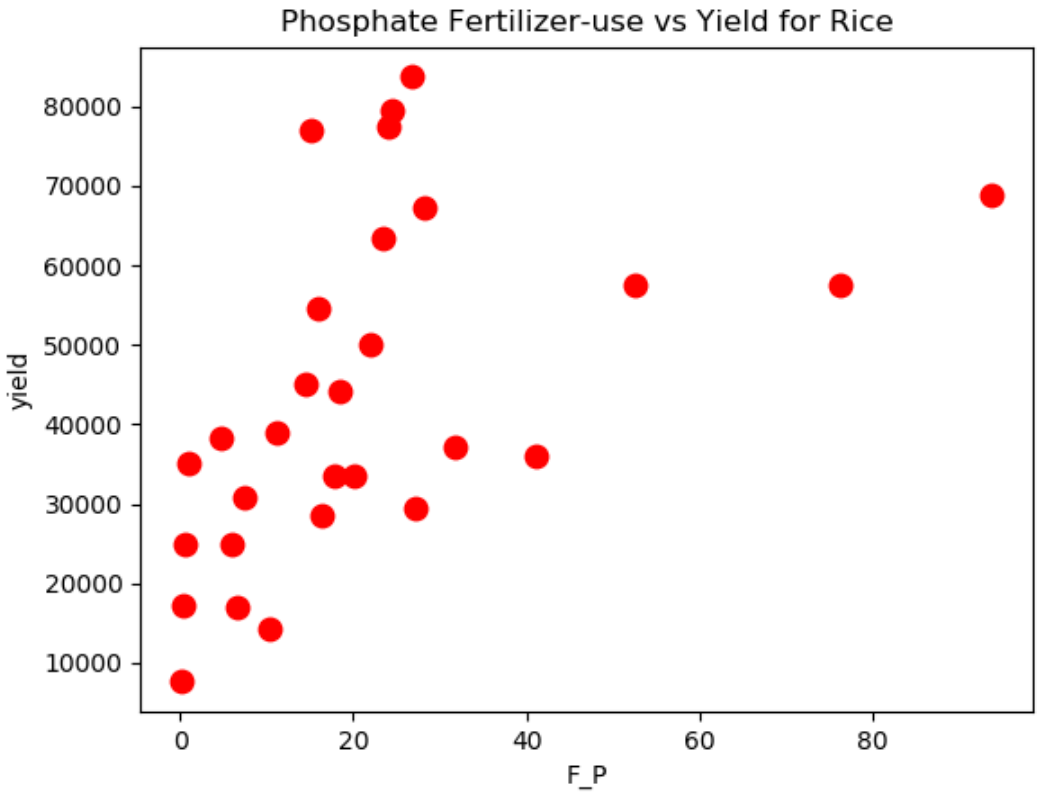
X-axis is % of farms with size between 5 and10 hectares. Most countries have no farm greater than 100 hectares.

Countries with high % of fields of size 5-10 hectares have good yield for Onion on an average.



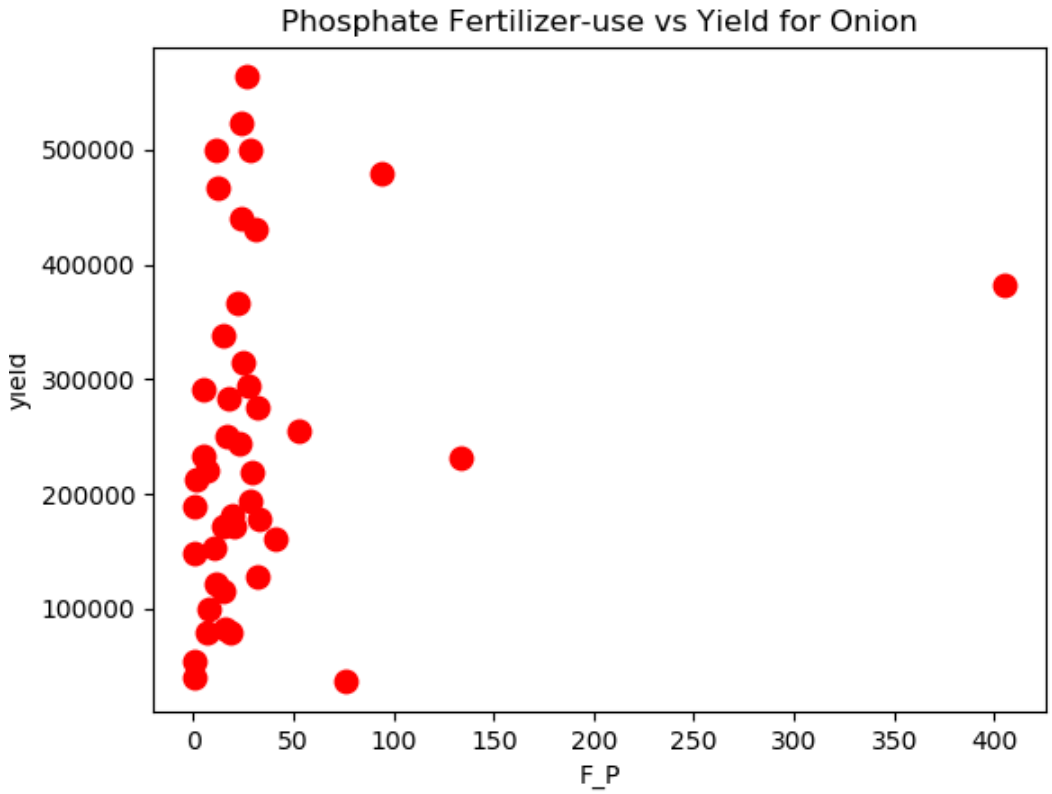
X-axis is kg/hectare of nitrogen fertilizer used.

Countries using moderate to high amount of nitrogen fertilizer per hectare tend to have better yield for wheat.



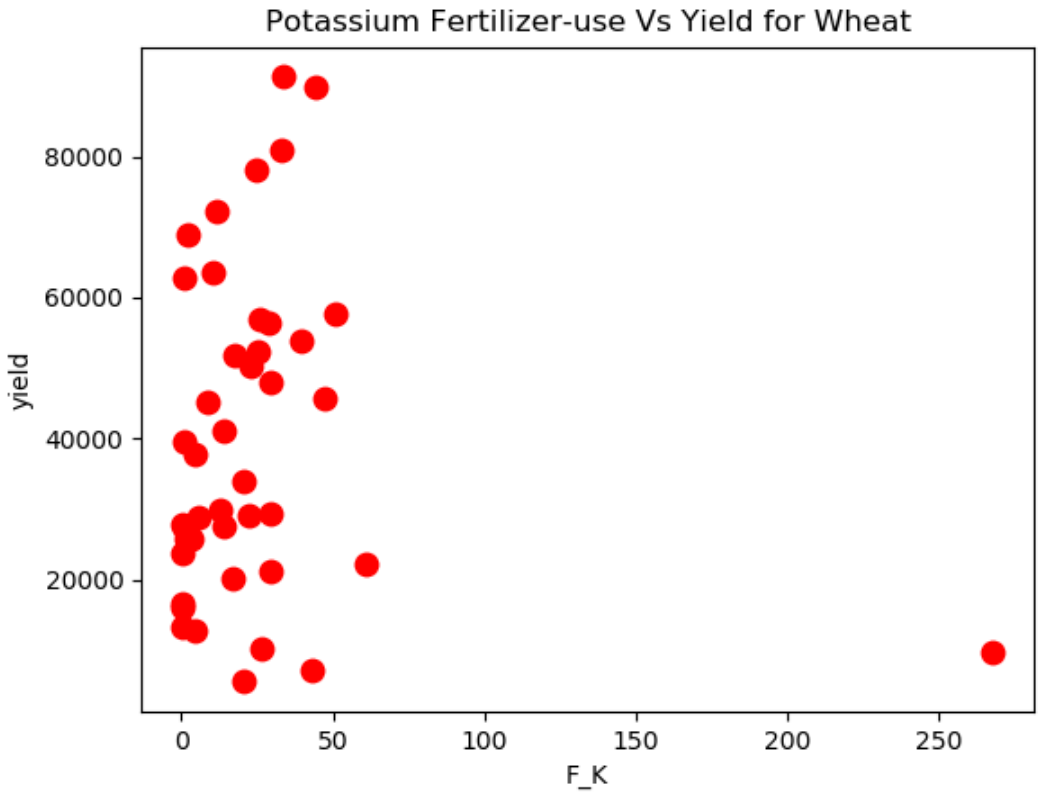
X-axis is kg/hectare of phosphate fertilizer used.

Countries that use moderate to high amount of phosphate fertilizer tend to have better yield for rice.



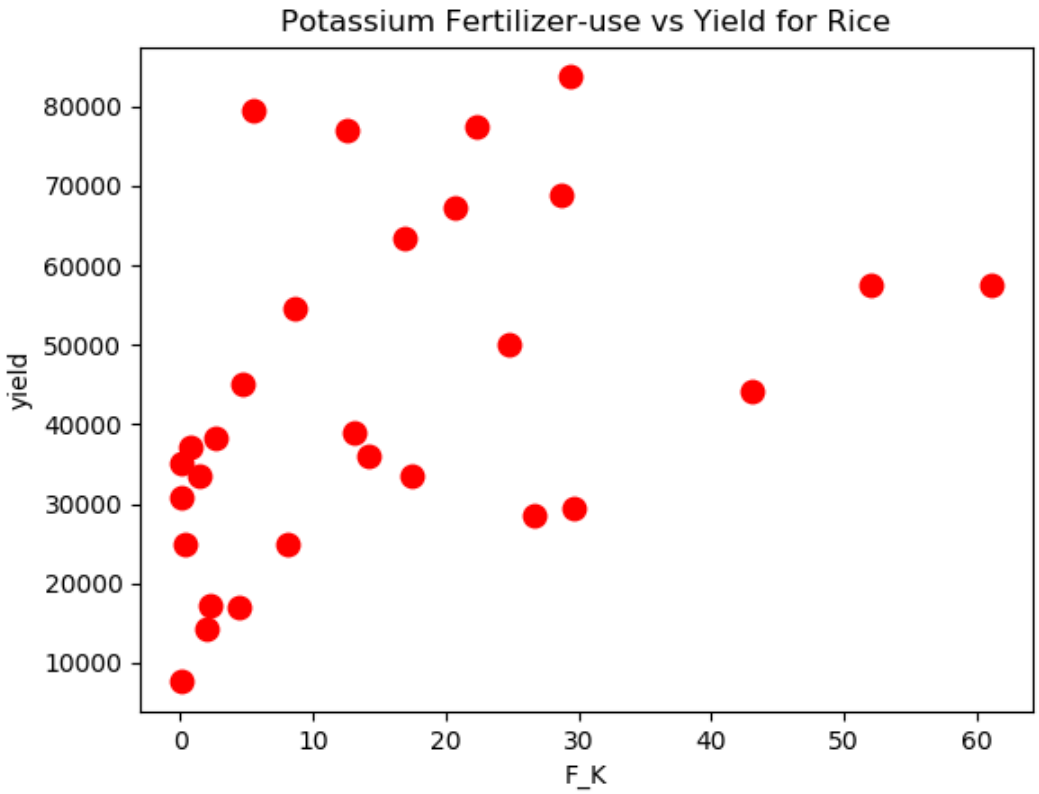
X-axis is kg/hectare of phosphate fertilizer used.

Countries that use small to moderate amount of phosphate have better yield for onions.



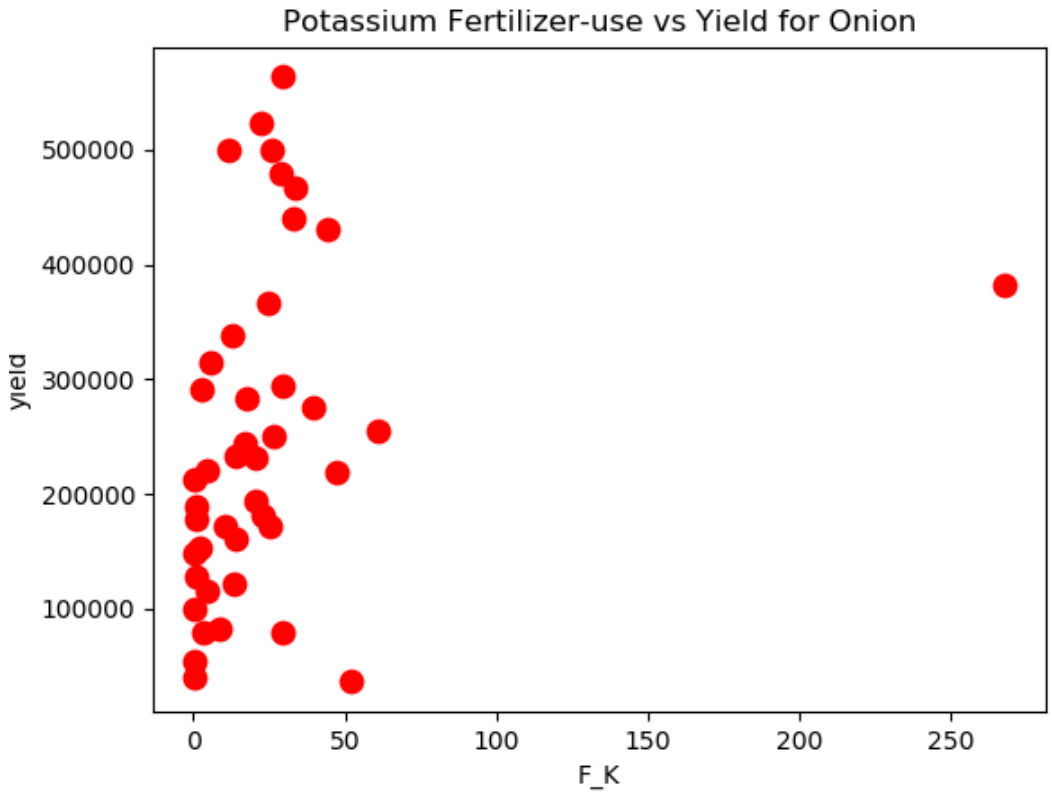
X-axis is kg/hectare of potassium fertilizer used.

Countries that use moderate amount of potassium fertilizer tend to have better yield for wheat.



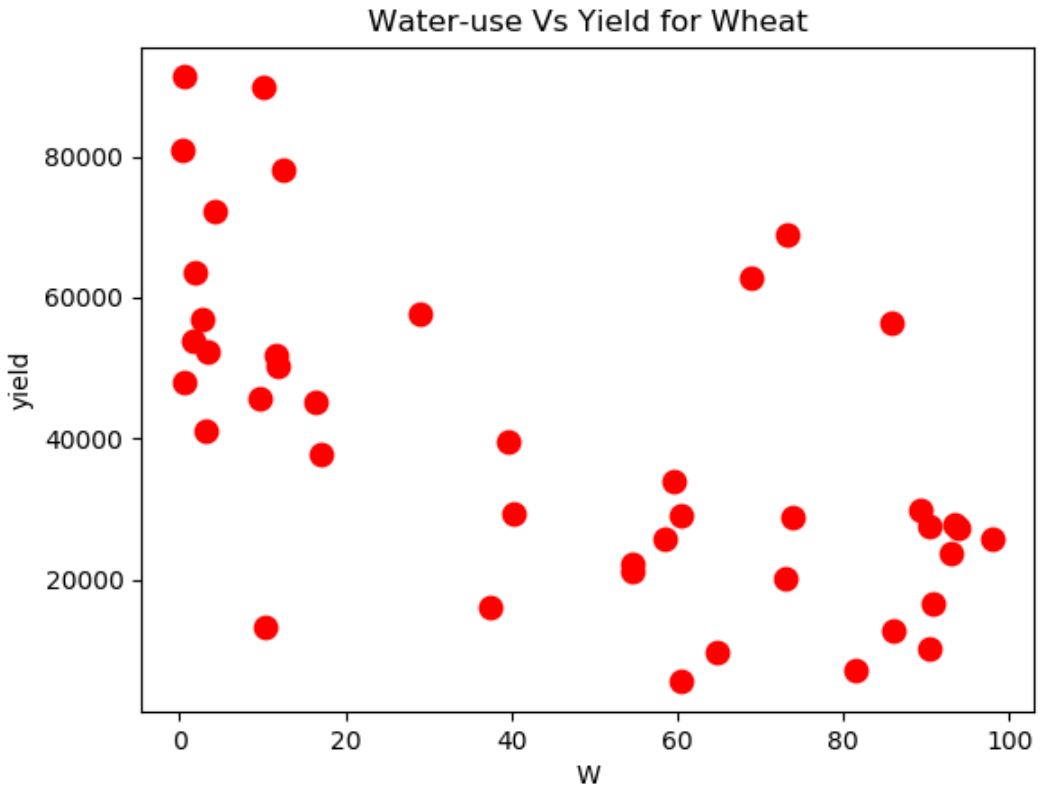
X-axis is kg/hectare of potassium fertilizer used.

Countries that use more potassium fertilizer in general tend to have better yield for rice.



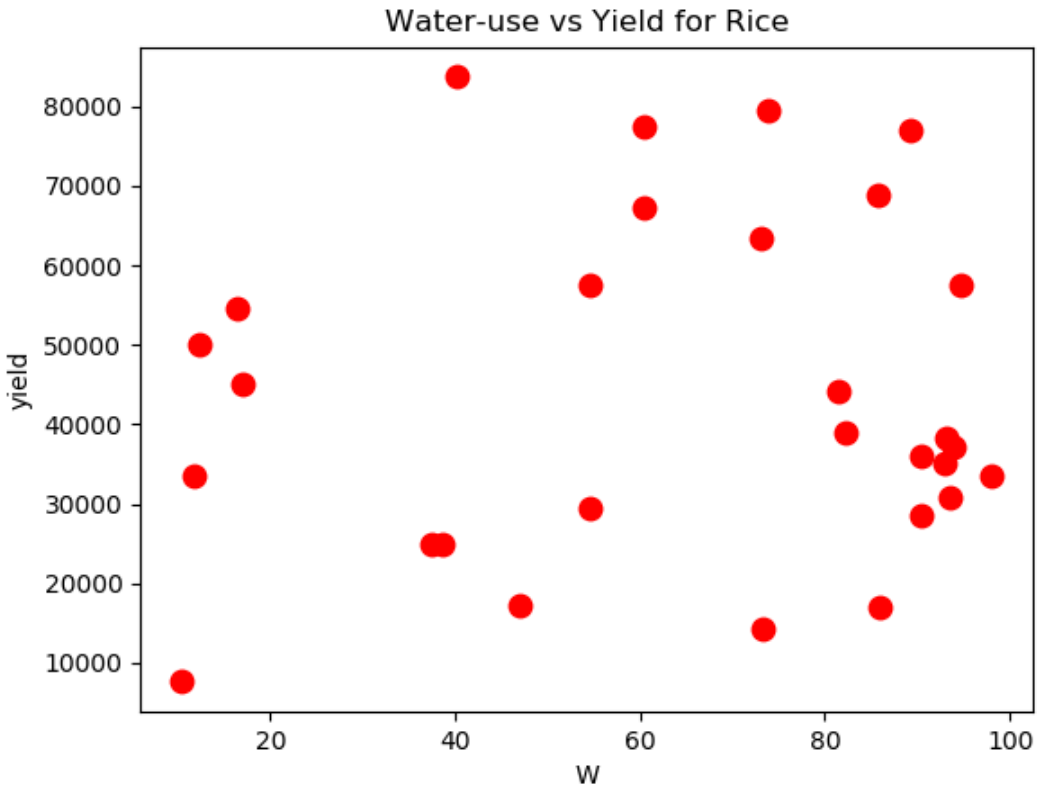
X-axis is kg/hectare of potassium fertilizer used.

Countries that use high amount of potassium fertilizer tend to have better yield for onion.



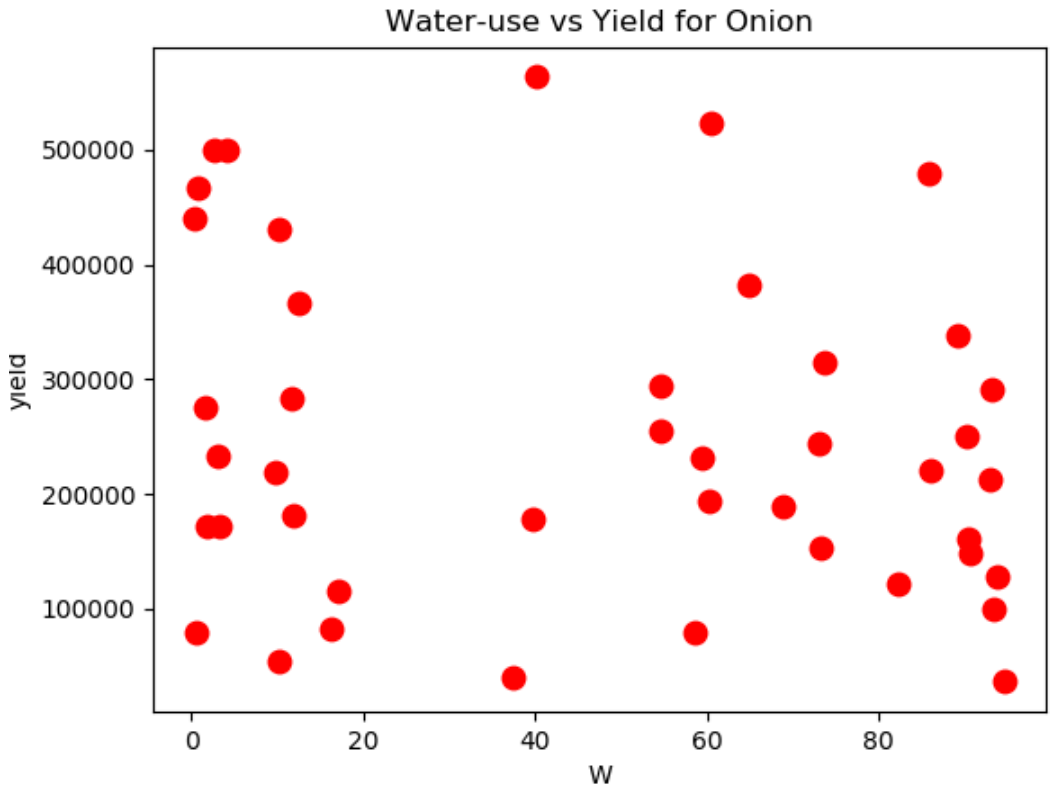
X-axis is % of total water used in agriculture.

Countries that use high amount of water not necessarily have better yield for wheat.



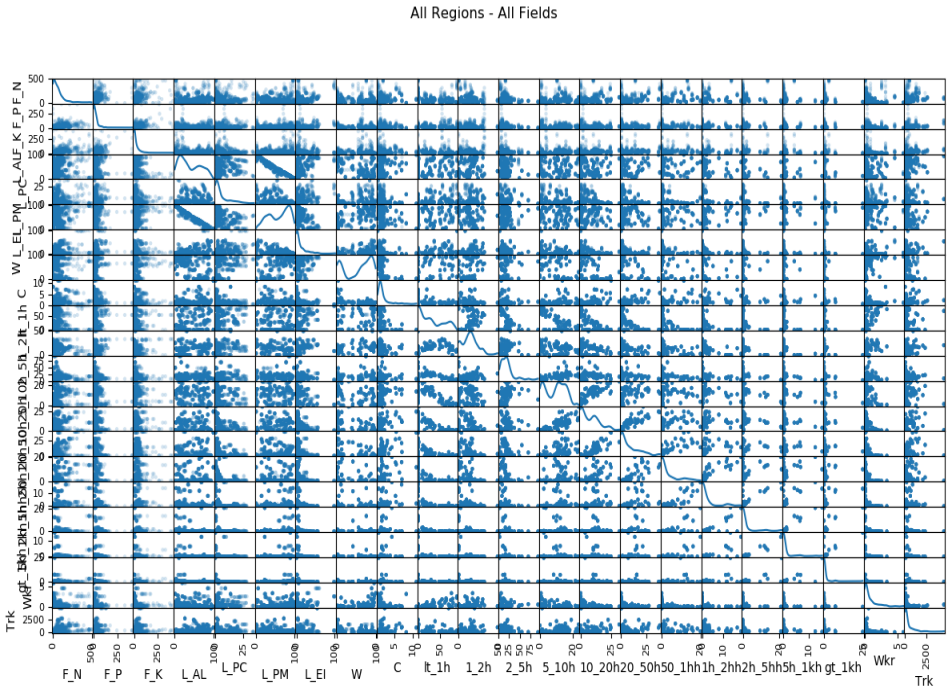
X-axis is % of total water used in agriculture.

Countries that use high amount of water tend to have better yield for rice.



No correlation between countries that use high amount of water and yield for onion.

Scatter matrix indicates negative correlation between amount of arable land (L_AL) and permanent meadows (L_PM). Countries that have more arable land have lesser permanent meadows. I will eliminate L_PM field before building our model.



Algorithms and Techniques

I use two algorithms for predicting yield:

MLR: Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable. All the variables for this study (water use, field size, fertilizer use etc.) are numeric values in the continuous range. So is the response variable Yield. Since yield for the three crops used vary in range, I create one model for each of the crops.

Random Forest Regression: This modeling technique operates by creating multiple decision trees and selecting the predicted value that is the mode of all values predicted. This technique avoids overfitting and generally provides better fit than multiple linear regression.

Both models do not need the input variables to be scaled or normalized.

Benchmark

This reference study http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0156571 gets a RMSE which is 11.9% of average observed yield for wheat, and a Pearson's r of 0.98. This reference study used climate data and other variables that my study does not consider.

III. Methodology

Data Preprocessing

In order to prepare for this project, data was downloaded from two sources:

- Food And Agriculture Organization of the United Nations (*http://www.fao.org/faostat/en/#data*)
- Nation Master (*http://www.nationmaster.com/country-info/groups/High-income-OECD-countries/Agriculture*)

Data Cleaning

The data from the above-mentioned sources is available in csv files. The data from both the sources was processed and consolidated. The next step was to eliminate fields and countries for which we had insufficient data. The process has been described in an earlier section. On the pruned set of fields and countries the missing data was filled in with data from the nearest year for which data was available for the crop and country.

Elimination of correlated column based on data exploration

Next fields that were highly correlated (positively or negatively) were eliminated. In this case value for arable land was negatively correlated with value for permanent meadows. Country that had a high area of arable land had less area of permanent meadows. This can be clearly observed in the scatter matrix.

Pandas DataFrame was heavily used for this task.

Implementation

All raw downloaded data in csv format is saved in data folder. There is one file each for data about fertilizer used, water used, land size, soil quality, land type etc. I have one class to process each file in the data_proc folder.

The python code is divided into 5 modules:
prepare_data.py – consolidates the csv files with data for different attributes.
sanitize_data.py – eliminate columns and countries with insufficient data. Fill missing data.
data_exploration.py – plots all the charts for each of the variables
generate_train_test_sets.py – generates training and test sets for each of the three crops
data_analysis.py – this module builds the models and prints the results
main_app.py – runs all the tasks above in one go and produces resut.txt file with the results from the run
Data files:
data/agri_data_final.csv – is the curated data file for the final analysis
data/*.csv – raw data from various sources

Once data was cleaned, the final dataset was saved as csv file containing columns for:

yield, item_id, year_code, 'F_N', 'F_P', 'F_K', 'L_AL', 'L_PC', 'L_PM', 'L_EI', 'W', 'GDP', 'C', 'country_name', 'lt_1h', '1_2h', '2_5h', '5_10h', '10_20h', '20_50h', '50_1hh', '1h_2hh', '2h_5hh', '5h_1kh', 'gt_1kh', 'Wkr', 'Trk'

This is the data on fertilizer used, land area for various kind as percentage of total agricultural area (arable land, permanent crop, permanent meadow, equipped for irrigation), water used, % of farms of various sizes, number of farm workers and number of tractors used per hectare of land.

Data for each of the three crops was split into training and test sets. Multiple linear regression model was fit to the training data and evaluated against the test data. The same was done with random forest regression.

Refinement

RMSE and R2 for Multiple Linear Regression models

Wheat			
RMSE Test as % of Average Yield	**RMSE Train as % of average Yield**	**R2-Test**	**R2-Train**
29.05	27.18	0.77	0.79
Rice Paddy			
RMSE Test as % of Average Yield	**RMSE Train as % of average Yield**	**R2-Test**	**R2-Train**
23.55	23.19	0.77	0.78
Onion			
RMSE Test as % of Average Yield	**RMSE Train as % of average Yield**	**R2-Test**	**R2-Train**
31.97	31.7	0.70	0.71

Random Forest regression with different parameters:

For random forest regression model I varied the number of trees built and criterion (loss function). The loss functions used were mean squared error (mse) and mean absolute error (mae). Here are the results for yield of wheat, rice paddy and onion. Highlighted rows are the parameter values for which we got best results on test sets for the criteria selected.

Wheat					
Number of Trees in the Random Forest Reg.	**Criterion**	**RMSE Test as % of Average Yield**	**RMSE Train as % of Average Yield**	**r-squared Test**	**r-squared Train**
6	mse	12.29	6.84	0.96	0.99
8	mse	11.77	6.97	0.96	0.99
10	mse	12.27	6.22	0.96	0.99
12	mse	11.90	7.41	0.96	0.99
16	mse	11.77	7.39	0.96	0.99
20	mse	12.72	6.97	0.96	0.99
25	mse	12.23	6.89	0.96	0.99
30	mse	12.38	6.74	0.96	0.99
35	mse	11.43	8.02	0.96	0.98

40	mse	12.00	6.80	0.96	0.99
6	mae	13.37	7.22	0.95	0.99
8	mae	12.70	7.03	0.96	0.99
10	mae	13.04	6.85	0.95	0.99
12	mae	12.66	7.50	0.96	0.98
16	mae	12.61	6.80	0.96	0.99
20	mae	12.77	6.91	0.95	0.99
25	mae	12.54	6.93	0.96	0.99
30	mae	12.12	6.29	0.96	0.99
35	mae	12.19	7.22	0.96	0.99
40	mae	13.55	6.86	0.95	0.99
Rice Paddy					
Number of Trees in the Random Forest Reg.	**Criterion**	**RMSE Test as % of Average Yield**	**RMSE Train as % of Average Yield**	**r-squared Test**	**r-squared Train**
6	mse	9.24	4.86	0.96	0.99
8	mse	9.56	4.91	0.96	0.99
10	mse	9.36	5.10	0.96	0.99
12	mse	11.00	4.90	0.95	0.99
16	mse	10.04	4.49	0.96	0.99
20	mse	9.98	5.19	0.96	0.99
25	mse	10.00	4.67	0.96	0.99
30	mse	10.59	4.46	0.95	0.99
35	mse	9.84	4.56	0.96	0.99
40	mse	10.22	4.55	0.96	0.99
6	mae	10.22	4.72	0.96	0.99
8	mae	11.32	4.72	0.95	0.99
10	mae	9.44	5.26	0.96	0.99
12	mae	10.06	4.52	0.96	0.99
16	mae	9.29	4.51	0.96	0.99
20	mae	9.99	4.62	0.96	0.99
25	mae	10.04	4.84	0.96	0.99
30	mae	10.32	5.09	0.96	0.99
35	mae	10.54	4.86	0.95	0.99
40	mae	10.24	4.87	0.96	0.99
Onion					
Number of Trees in the Random	**Criterion**	**RMSE Test as % of Average Yield**	**RMSE Train as % of Average Yield**	**r-squared Test**	**r-squared Train**

Forest Reg.					
6	mse	16.79	6.78	0.92	0.99
8	mse	16.57	6.84	0.92	0.99
10	mse	16.83	6.52	0.92	0.99
12	mse	16.63	6.33	0.92	0.99
16	mse	17.31	6.85	0.91	0.99
20	mse	17.24	6.90	0.91	0.99
25	mse	16.77	6.03	0.92	0.99
30	mse	17.58	6.27	0.91	0.99
35	mse	17.37	6.71	0.91	0.99
40	mse	17.17	7.15	0.91	0.99
6	mae	16.97	6.28	0.92	0.99
8	mae	17.20	6.56	0.91	0.99
10	mae	17.63	7.22	0.91	0.99
12	mae	17.17	6.88	0.91	0.99
16	mae	17.48	7.96	0.91	0.99
20	mae	17.99	6.61	0.91	0.99
25	mae	17.02	7.08	0.92	0.99
30	mae	17.06	6.62	0.92	0.99
35	mae	17.31	6.62	0.91	0.99
40	mae	19.90	7.05	0.91	0.99

IV. Results

Model Evaluation and Validation

Random Forest Regression model provided considerably lower RMSE and higher r-squared values compared to the Multiple Regression Model. The model was trained on training set and evaluated on test set. The various values for the parameters did not change the result of RMSE and r-squared values significantly.

Results for yield of wheat:

- For yield of wheat RMSE of around 11.5% of average yield was achieved on test set with r-squared value of 0.96.
- The input variables that influenced the yield of wheat in decreasing order: ['gt_1hh', '2_5h', '10_20h', '5_10h', 'F_K', '50_1hh', 'Trk', 'F_P', 'L_AL', 'L_EI', 'C', 'lt_1h', '1_2h', 'Wkr', 'W', 'L_PC', 'F_N', '20_50h'].

This means countries with large farm size (greater than 100 hectares) produced the highest yield. The second variable that influenced the yield the most was percentage of farms with sizes of 2-5 hectares.

Result for yield of rice paddy:

- For yield of rice paddy RMSE was around 9.92% of average yield on test set with r-squared value of 0.96.
- The input variables that influenced the yield for rice paddy in decreasing order:
  ['50_1hh', 'Wkr', '5_10h', '20_50h', '2_5h', 'C', '10_20h', 'gt_1hh', 'lt_1h', 'W', 'F_P', 'F_K', 'L_AL', 'L_PC', '1_2h', 'L_EI', 'F_N', 'Trk'] .
  This means countries with high % of farm of sizes between 50 and 100 hectares produce higher yield for rice. Also countries that employ more labor per hectare do well with yield.

Result for yield of onion:

- For yield of onion RMSE was around 16.5% of average yield on test set with r-squared value of 0.92. The reason for higher errors in the model could be due to high variation in yield of onion for the high-income countries (see box plot for yield in data exploration section) which make up a third of the data we considered.
- The input variables that influenced the yield of onion in decreasing order:
  ['10_20h', '50_1hh', '2_5h', 'gt_1hh', 'L_EI', 'Trk', '5_10h', 'F_K', 'F_P', 'W', 'L_AL', 'C', 'F_N', '1_2h', 'Wkr', 'lt_1h', 'L_PC', '20_50h']
  Here again medium to small farm sizes produce better yield.

Justification

Random forest regression model gives us better results compared to multiple linear regression. Predicting yield of crops is a complicated problem to solve. In this project I have not considered climate conditions, seed quality etc. that influence the yield considerably. The yield could also be impacted by availability of water for irrigation in particular regions during a particular year. Generally these kinds of studies are done for small geographical regions (district level) with input of climate data, soil condition, water usage etc. for specific crops of interest. This study only took country averages of water usage, farm size, fertilizer used etc. and applied it uniformly to all crops. This is not ideal approach for understanding this problem.

I cannot justify that this model truly captures the real world situation. The result is as good as the input data I have. The input data curated from FAO of United Nations and other sources are not nearly sufficient enough to understand this issue of yield of crops.

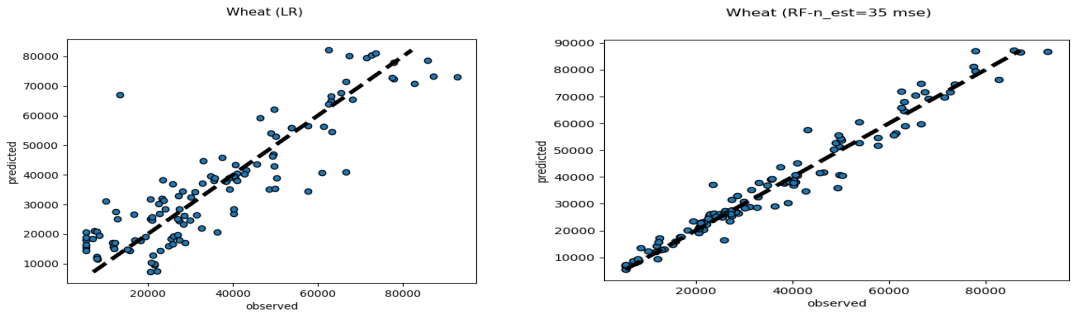
Having said all that, for the data that I had available, I believe this random forest regression model is a good choice.

V. Conclusion

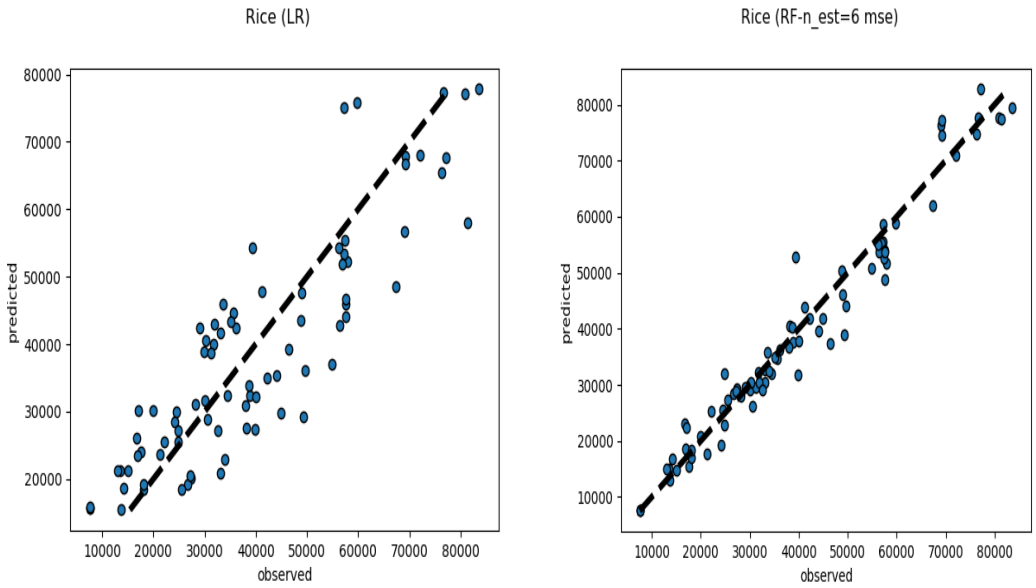
Free-Form Visualization

Plots of actual yield vs predicted yield by the model is a good indicator of how well the regression fit was.

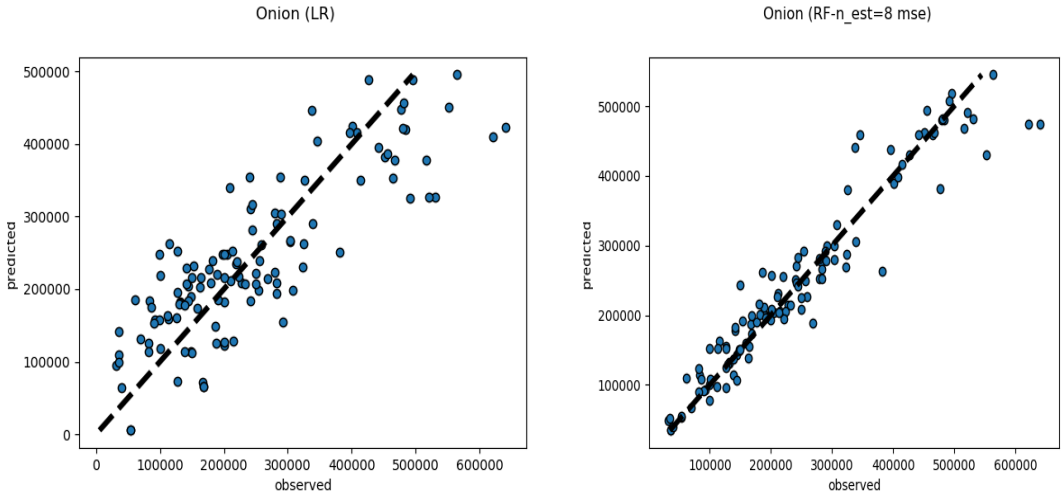
Multiple linear regression vs random forest fit for yield of Wheat:



Multiple linear regression vs Random forest for yield of rice paddy:



Multiple linear regression vs Random forest fit for yield of onion:



In all these plots it is quite evident that Random Forest Regression provides lesser error in predicting yield.

Reflection

Much of the time was spent in evaluating agriculture data from various sources. Evaluating the quality of data and making decisions on what data to use and what to throw away consumed lot of time. The work done here will not help us understand the real world problem of yield of crops but with more relevant data and applying this to a smaller geographic region would be the next step towards refining this algorithm.

Improvement

I would have liked to have more data available at granularity of crop type rather than applying country averages to all crops studied here. Climate data and a time-series like approach to track yield over a period of time would be useful. I have data for different years but I did not treat those as a time-series and rather treated it as just another data point for the algorithm. There are other related questions about sustainable farming, organic farming, waste reduction in farming etc. Lot more work needs to be done in this area. I would like to study more research papers and find out what the challenges are and how people have tried to understand it better. Satellite images could be used to provide climate and soli conditions. Overall the exercise of working with raw data from scratch and not using curated data (like in Kaggle projects) provided me with a better feel for real world projects.