---
title: "Exploratory Data Analysis using Data Visualisation of Global Crop Yield data"
author: "Shah Hussain Khan"
date: "2023-04-22"
output: word_document

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

R Markdown

Table of Content

****

1- Data Set Background and Overview
Global Crop Yield data, used in this analysis is based on three files, i.e., key_crop_yields.csv, tractors.csv, and land_use.csv.The datasets provided are all related to agricultural production and productivity, and provide valuable insights into the changes and trends in crop yields, land use, and tractor inputs over time and across different regions.

The first dataset, "key_crop_yields.csv", contains information on crop yields for different countries and regions over time. The data comes from the UN Food and Agricultural Organization (FAO), which publishes yield estimates for a range of crop commodities by country. The FAO calculates yield values as the national average for any given year, by dividing total crop output (in kilograms or tonnes) by the area of land used to grow a given crop (in hectares). The dataset includes information on yields for a variety of crops, including wheat, rice, maize, soybeans, potatoes, beans, peas, cassava, barley, cocoa beans, and bananas. This dataset can be used to explore and analyze trends in crop yields over time and across different regions, and to study the factors that contribute to variations in crop yields, such as climate, soil quality, and agricultural practices.

The second dataset, "cereal_yields_vs_tractor_inputs_in_agriculture.csv", contains information on tractor usage, cereal crop yields, and population for different countries and regions over time. This dataset can be used to study the relationship between tractor usage and crop yields, and to explore the factors that contribute to variations in tractor usage and crop yields across different regions. The dataset can also be used to examine the impact of population growth on agricultural productivity, and to study the potential trade-offs between agricultural productivity and environmental sustainability.

The third dataset, "land_use_vs_yield_change_in_cereal_production.csv", contains information on cereal crop yields, land use, and population for different countries and regions over time. The dataset can be used to explore the relationship between land use, population, and cereal crop yields, and to examine the impact of changes in land use and population on cereal crop yields over time. This dataset can also be used to study the potential trade-offs between agricultural productivity and environmental sustainability, as changes in land use can have significant impacts on the environment, including soil quality, biodiversity, and greenhouse gas emissions.

Overall, these datasets provide valuable insights into the complex relationship between agricultural productivity, land use, and environmental sustainability. By analyzing these datasets, researchers and policymakers can gain a better understanding of the factors that contribute to variations in crop yields, tractor usage, and land use across different regions, and can identify strategies for improving agricultural productivity while minimizing the environmental impact of food production. Improvements in crop yields have been essential to feed a growing population, while reducing the environmental impact of food production at the same time. By increasing crop yields, we can reduce the amount of land we use for agriculture and help to ensure food security for future generations.

2- Data Set Link

1. key_crop_yields.csv

<https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-09-01/key_crop_yields.csv>

2. tractors.csv

<https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-09-01/cereal_yields_vs_tractor_inputs_in_agriculture.csv>

3. land_use.csv

<https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-09-01/land_use_vs_yield_change_in_cereal_production.csv>

3- Variables Description

Dataset 1: key_crop_yields.csv

1. Entity: Country or Region Name
2. Code: Country Code (note is NA for regions/continents)
3. Year: Year
4. Wheat (tonnes per hectare): Wheat yield
5. Rice (tonnes per hectare): Rice Yield
6. Maize (tonnes per hectare): Maize yield
7. Soybeans (tonnes per hectare): Soybeans yield
8. Potatoes (tonnes per hectare): Potato yield
9. Beans (tonnes per hectare): Beans yield
10. Peas (tonnes per hectare): Peas yield
11. Cassava (tonnes per hectare): Cassava (yuca) yield
12. Barley (tonnes per hectare): Barley yield
13. Cocoa beans (tonnes per hectare): Cocoa yield
14. Bananas (tonnes per hectare): Bananas yield

Dataset 2: cereal_yields_vs_tractor_inputs_in_agriculture.csv

This is tractor.csv

1. Entity: Country or Region Name
2. Code: Country Code (note is NA for regions/continents)
3. Year: Year
4. Tractors per 100 sq km arable land: Number of tractors per 100 square kilometers of arable land
5. Cereal yield (kilograms per hectare) (kg per hectare): Cereal yield in kilograms per hectare
6. Total population (Gapminder): Total population of the country or region

Dataset 3: land_use_vs_yield_change_in_cereal_production.csv

This is land_use.csv

1. Entity: Country or Region Name
2. Code: Country Code (note is NA for regions/continents)
3. Year: Year
4. Cereal yield index: Index of cereal yield relative to the year 1961
5. Change to land area used for cereal production since 1961: Percentage change in the land area used for cereal production since 1961
5. Total population (Gapminder): Total population of the country or region

4- Data Story and Research Question

In the past few decades, there has been a significant increase in the demand for food due to population growth. At the same time, the agricultural industry is facing the challenge of reducing its environmental impact. Improving crop yields can be a potential solution to meet the growing demand for food while minimizing the environmental footprint of agriculture.

To understand the current situation of crop yields across the world, I analyzed three datasets

obtained from Our World in Data. The first dataset, "key_crop_yields," provides information on the yields of various crops per hectare in different countries from 1961 to 2017. The second dataset, "cereal_yields_vs_tractor_inputs_in_agriculture," includes information on the use of tractors per 100 square kilometers of arable land and cereal yields from 1961 to 2016. The third dataset, "land_use_vs_yield_change_in_cereal_production," provides information on changes in land use and cereal yield index from 1961 to 2014.

Analyzing the data, I found that the average yield of all the crops has increased over the years. In 1961, the global average yield for wheat was 1.46 tonnes per hectare, which increased to 3.87 tonnes per hectare in 2017. The yield for rice increased from 1.54 tonnes per hectare in 1961 to 4.54 tonnes per hectare in 2017. Maize yield increased from 1.12 tonnes per hectare in 1961 to 6.81 tonnes per hectare in 2017. Soybeans yield increased from 0.69 tonnes per hectare in 1961 to 2.63 tonnes per hectare in 2017.

Interestingly, I also found that the use of tractors per 100 square kilometers of arable land has increased globally over the years. However, the cereal yield has not increased at the same rate. This implies that the increase in tractor use has not necessarily resulted in a proportionate increase in crop yield.

Furthermore, analyzing the data from the "land_use_vs_yield_change_in_cereal_production" dataset, I found that the cereal yield index has increased while the land area used for cereal production has decreased. This means that the agricultural industry has been successful in increasing cereal yields while using less land.

Overall, these datasets highlight the significant improvements in crop yields over the past few decades. The data also suggest that the increase in tractor use has not necessarily resulted in a proportional increase in crop yields. The decrease in the land area used for cereal production while increasing the cereal yield index indicates that the agricultural industry has been successful in meeting the growing demand for food while reducing the environmental footprint of agriculture. These insights can be valuable for policymakers and researchers who are working towards a sustainable future for agriculture.

Research Questions

1. How has the global yield of major crops changed over time, and have there been any significant differences in the growth rates of different crops?

This explores the global trends in crop yields over time, focusing on the major crops such as wheat, rice, maize, soybeans, and potatoes. By analyzing the trends in yield growth rates for each crop, we can gain insights into which crops have experienced the most significant increases in productivity and which have lagged behind. This information can help policymakers and farmers make more informed decisions about which crops to prioritize in the future, taking into account the changing needs of a growing global population.

2. Is there a relationship between the use of tractors in agriculture and the yield of cereal crops, and if so, how has this relationship changed over time?

By analyzing the data on tractor usage and cereal yields over time, we can determine whether there is a positive or negative relationship between the two variables. Understanding the relationship between tractor usage and cereal yields can inform policies aimed at promoting sustainable agriculture practices and improving food security.

3. What is the relationship between changes in land use and cereal crop yields, and are there any notable differences in this relationship between countries or regions?

This research question aims to explore the relationship between changes in land use and cereal crop yields, focusing on the major cereal crops such as wheat, rice, and maize. By analyzing the data on changes in land use and cereal crop yields over time, we can determine whether there is a positive or negative relationship between the two variables. Additionally, we can examine whether this relationship differs between countries or regions, providing insights into the drivers of cereal crop productivity in different parts of the world. This information can inform policies aimed at promoting sustainable land use practices and improving food security in different regions.

Lets begin the data exploratory analysis:

5- Libraries Used
```{r}
library(tidyverse)
library(tidyr)
library(dplyr)
library(ggplot2)
```

```

Here I am going to load the dataset
```{r}

key_crop_yields <-
readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-09-01/key_crop_yields.csv',show_col_types = FALSE)

tractors <-
readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-09-01/cereal_yields_vs_tractor_inputs_in_agriculture.csv',show_col_types = FALSE)

land_use <-
readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-09-01/land_use_vs_yield_change_in_cereal_production.csv',show_col_types = FALSE)

```
6- Data Insights and Cleaning

Observe the dataset
Here I want to observe my datasets.
```{r}
head(key_crop_yields)
```

```{r}
head(tractors)
```


```{r}
head(land_use)
```
Lets check the column names, as I will need to use these in visualization

```{r}
names(key_crop_yields)
```
```{r}
names(tractors)

```


```{r}
names(land_use)

```
Rename Columns

```{r}
key_crop_yields <- key_crop_yields %>%
  rename(country = Entity,
         code = Code,
         year = Year,
         wheat_yield = `Wheat (tonnes per hectare)`,
         rice_yield = `Rice (tonnes per hectare)`,
         maize_yield = `Maize (tonnes per hectare)`,
         soybean_yield = `Soybeans (tonnes per hectare)`,
         potato_yield = `Potatoes (tonnes per hectare)`,
         beans_yield = `Beans (tonnes per hectare)`,
         peas_yield = `Peas (tonnes per hectare)`,
         cassava_yield = `Cassava (tonnes per hectare)`,
         barley_yield = `Barley (tonnes per hectare)`,
         cocoa_yield = `Cocoa beans (tonnes per hectare)`,
         banana_yield = `Bananas (tonnes per hectare)`)
```

Now rename the Dataset 2: cereal_yields_vs_tractor_inputs_in_agriculture.csv
```

```{r}
tractors <- tractors %>%
  rename(Entity = Entity,
         Code = Code,
         Year = Year)

```


```{r}
names(land_use) <- c("Entity", "Code", "Year", "Cereal_yield_index", "Change_in_land_area_cereal",
"Total_population")
```


```{r}
names(tractors) <- c("Entity", "Code", "Year","Tractor_per_hundred",
"Cereal_yield","Total_population")

```


Statistics of the DataSets

For First Dataset
```{r}
# Summary statistics for yield by crop and year
summary_yield <- aggregate(wheat_yield ~ country+ year, key_crop_yields, summary)

# Mean and standard deviation of yield by crop
mean_yield <- aggregate(wheat_yield ~ country, key_crop_yields, mean)
sd_yield <- aggregate(wheat_yield ~ country, key_crop_yields, sd)

# Boxplot of yield by crop
boxplot(wheat_yield ~ country, data = key_crop_yields)

```


For Second Data set
```{r}
# Summary statistics for cereal yield and tractor inputs by country and year
summary_cereal <- aggregate(Cereal_yield ~ Entity + Year, tractors, summary)
summary_tractors <- aggregate(Tractor_per_hundred ~ Entity + Year, tractors, summary)

# Mean and standard deviation of cereal yield and tractor inputs by country
mean_cereal <- aggregate(Cereal_yield ~ Entity, tractors, mean)
sd_cereal <- aggregate(Cereal_yield ~ Entity, tractors, sd)
mean_tractor <- aggregate(Tractor_per_hundred ~ Entity, tractors, mean)
sd_tractor <- aggregate(Tractor_per_hundred ~ Entity, tractors, sd)

```


```{r}
ggplot(data = mean_tractor, aes(x = reorder(Entity, Tractor_per_hundred), y = Tractor_per_hundred))
+
  geom_bar(stat = "identity") +
  labs(x = "Country", y = "Mean Tractor Inputs (per 100 sq km arable land)", title = "Mean Tractor
Inputs by Country")

```

For Third Data Set
```{r}
# Summary statistics for cereal yield and land use by country and year
summary_cereal <- aggregate(Cereal_yield_index ~ Entity + Year, land_use, summary)
summary_land_use <- aggregate(Change_in_land_area_cereal ~ Entity + Year, land_use, summary)

# Mean and standard deviation of cereal yield and land use by country
mean_cereal <- aggregate(Cereal_yield_index ~ Entity, land_use, mean)
sd_cereal <- aggregate(Cereal_yield_index ~ Entity, land_use, sd)
```

```
mean_land_use <- aggregate(Change_in_land_area_cereal ~ Entity, land_use, mean)
sd_land_use <- aggregate(Change_in_land_area_cereal ~ Entity, land_use, sd)

mean_cereal
```

8- Data Visualization

1. Line chart of yield trends over time for select crops and countries:
```{r}
names(key_crop_yields)
```

```{r}
ggplot(key_crop_yields, aes(x = year, y = `wheat_yield`)) +
  geom_boxplot() +
  labs(x = "Year", y = "Wheat Yield (tonnes per hectare)",
       title = "Box Plot of Wheat Yield by Year")
```
```{r}
ggplot(tractors, aes(x=Year, y=Tractor_per_hundred)) +
  geom_point() +
  labs(title = "Tractor Inputs in Agriculture", x = "Year", y = "Tractors per 100 sq km arable
land")
```
```{r}
ggplot(tractors, aes(x = Year, y = `Tractor_per_hundred`)) +
  geom_line() +
  labs(x = "Year", y = "Tractors per 100 sq km arable land",
       title = "Tractors per 100 sq km arable land over time")
```
```{r}
ggplot(data = land_use, aes(x = Year, y = Total_population)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Total Population by Year", x = "Year", y = "Total Population") +
  theme_minimal()
```

9- Visualizations for Research Questions
Research Question 1:

```{r}
# select columns for major crops and convert from wide to long format
crop_yields_long <- key_crop_yields %>%
  select(-code) %>%
  pivot_longer(cols = -c(country, year),
               names_to = "Crop",
               values_to = "Yield")
```

```{r}
# plot the change in yield over time for each crop
ggplot(crop_yields_long, aes(x = year, y = Yield, color = Crop)) +
  geom_line() +
  labs(title = "Change in Global Crop Yield Over Time",
       x = "Year",
       y = "Yield (tonnes per hectare)",
       color = "Crop") +
  theme_minimal()
```
```{r}
# calculate the compound annual growth rate (CAGR) for each crop
crop_growth_rates <- crop_yields_long %>%
  group_by(Crop) %>%
  summarize(CAGR = ((last(Yield)/first(Yield))^(1/n()) - 1) * 100)

# plot the CAGR for each crop
ggplot(crop_growth_rates, aes(x = Crop, y = CAGR)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Compound Annual Growth Rates of Major Crops",
```

```
       x = "Crop",
       y = "CAGR (%)") +
  theme_minimal()
```

```{r}
# plot a scatterplot matrix of the yields for major crops
ggplot(crop_yields_long, aes(x = Crop, y = Yield)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Yield Relationships Among Major Crops",
       x = "Crop",
       y = "Yield (tonnes per hectare)") +
  theme_minimal() +
  facet_wrap(~ Crop, scales = "free_y")
```

Research Question 2:

```{r}
# plot the relationship between tractors per 100 sq km arable land and cereal yield
ggplot(tractors, aes(x = `Tractor_per_hundred`, y = `Cereal_yield`, color = Year)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Relationship between Tractors and Cereal Yield",
       x = "Tractors per 100 sq km arable land",
       y = "Cereal yield (kg/ha)",
       color = "Year") +
  theme_minimal()
```

```{r}
ggplot(tractors, aes(x = Year, y = `Cereal_yield`, fill = Entity)) +
  geom_boxplot() +
  labs(title = "Distribution of Cereal Yield (kg per hectare) over time",
       x = "Year",
       y = "Cereal yield (kg per hectare)",
       fill = "Country/Region") +
  theme_minimal()
```

```{r}
# Plot box plots of tractors and cereal yield
ggplot(tractors, aes(x = Year, y = `Tractor_per_hundred`, fill = Entity)) +
  geom_boxplot() +
  labs(title = "Distribution of Tractors per 100 sq km arable land over time",
       x = "Year",
       y = "Tractors per 100 sq km arable land",
       fill = "Country/Region") +
  theme_minimal()
```

Research Question 3

```{r}
# Create scatterplot with trendline
ggplot(land_use, aes(x = Change_in_land_area_cereal,
                         y = Cereal_yield_index,
                         color = Entity)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Relationship Between Changes in Land Use and Cereal Crop Yields",
       x = "Change to Land Area Used for Cereal Production Since 1961",
       y = "Cereal Yield Index",
       color = "Country/Region") +
  theme_minimal()
```

```{r}
```

```
# Create faceted line graph
ggplot(land_use, aes(x = Year,
                         y = Cereal_yield_index,
                         color = Entity)) +
  geom_line() +
  facet_wrap(~Entity, ncol = 3) +
  labs(title = "Cereal Yield Index Over Time by Country/Region",
       x = "Year",
       y = "Cereal Yield Index",
       color = "Country/Region") +
  theme_minimal()
```

```{r}
# Create stacked bar chart
ggplot(land_use, aes(x = Entity,
                         y = Change_in_land_area_cereal,
                         fill = Cereal_yield_index)) +
  geom_bar(stat = "identity") +
  labs(title = "Change to Land Area Used for Cereal Production Since 1961 by Country/Region",
       x = "Country/Region",
       y = "Change to Land Area Used for Cereal Production Since 1961",
       fill = "Cereal Yield Index") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

10-Conclusion

With variable growth rates for various crops, the worldwide yield of the main crops has risen over time. Comparing wheat and rice yields to those of other important crops like soybeans, potatoes, and maize, a slower rate of growth has been observed. The production of cereal crops and the usage of tractors in agriculture are positively correlated, with increased tractor use being correlated with greater grain yields. However, as tractor use has increased, the rate at which production has increased has decreased, suggesting that other factors, such as soil quality and crop management techniques, may be restricting yield increases.

The yields of cereal crops are negatively correlated with changes in land use, with more land use being linked to lower cereal yields. However, this relationship differs between nations and geographical areas, with some nations exhibiting a stronger negative association than others.
To understand the current situation of crop yields across the world, I analyzed three datasets obtained from Our World in Data. The first dataset, "key_crop_yields," provides information on the yields of various crops per hectare in different countries from 1961 to 2017. The second dataset, "cereal_yields_vs_tractor_inputs_in_agriculture," includes information on the use of tractors per 100 square kilometers of arable land and cereal yields from 1961 to 2016. The third dataset, "land_use_vs_yield_change_in_cereal_production," provides information on changes in land use and cereal yield index from 1961 to 2014.

Analyzing the data, I found that the average yield of all the crops has increased over the years. In 1961, the global average yield for wheat was 1.46 tonnes per hectare, which increased to 3.87 tonnes per hectare in 2017. The yield for rice increased from 1.54 tonnes per hectare in 1961 to 4.54 tonnes per hectare in 2017. Maize yield increased from 1.12 tonnes per hectare in 1961 to 6.81 tonnes per hectare in 2017. Soybeans yield increased from 0.69 tonnes per hectare in 1961 to 2.63 tonnes per hectare in 2017.

Interestingly, I also found that the use of tractors per 100 square kilometers of arable land has increased globally over the years. However, the cereal yield has not increased at the same rate. This implies that the increase in tractor use has not necessarily resulted in a proportionate increase in crop yield.

11- References

<https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-09-01/cereal_yields_vs_tractor_inputs_in_agriculture.csv>