Exploratory Data Analysis using Data Visualisation of Global Crop Yield data

Shah Hussain Khan

2023-04-22

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
6. 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

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.2 ──
## ✓ ggplot2 3.4.0      ✓ purrr   0.3.5
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.5.0
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(tidyr)
library(dplyr)
library(ggplot2)
```

Here I am going to load the dataset

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

tractors <-
readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesda
y/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/tidytuesda
y/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.

```
head(key_crop_yields)
```

```
## # A tibble: 6 × 14
##   Entity      Code   Year Wheat…¹ Rice …² Maize…³ Soybe…⁴ Potat…⁵ Beans…⁶
```

```
Peas …⁷
##   <chr>      <chr> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
<dbl>
## 1 Afghanist… AFG    1961   1.02     1.52    1.4       NA    8.67      NA
NA
## 2 Afghanist… AFG    1962   0.974    1.52    1.4       NA    7.67      NA
NA
## 3 Afghanist… AFG    1963   0.832    1.52    1.43      NA    8.13      NA
NA
## 4 Afghanist… AFG    1964   0.951    1.73    1.43      NA    8.6       NA
NA
## 5 Afghanist… AFG    1965   0.972    1.73    1.44      NA    8.8       NA
NA
## 6 Afghanist… AFG    1966   0.867    1.52    1.44      NA    9.07      NA
NA
## # … with 4 more variables: `Cassava (tonnes per hectare)` <dbl>,
## #   `Barley (tonnes per hectare)` <dbl>,
## #   `Cocoa beans (tonnes per hectare)` <dbl>,
## #   `Bananas (tonnes per hectare)` <dbl>, and abbreviated variable names
## #   ¹`Wheat (tonnes per hectare)`, ²`Rice (tonnes per hectare)`,
## #   ³`Maize (tonnes per hectare)`, ⁴`Soybeans (tonnes per hectare)`,
## #   ⁵`Potatoes (tonnes per hectare)`, ⁶`Beans (tonnes per hectare)`, …

head(tractors)

## # A tibble: 6 × 6
##   Entity      Code  Year  `Tractors per 100 sq km arable land` Cereal …¹
Total…²
##   <chr>       <chr> <chr>                                <dbl>     <dbl>
<dbl>
## 1 Afghanistan AFG   1961                                 0.157     1115.
9.17e6
## 2 Afghanistan AFG   1962                                 0.195     1079
9.35e6
## 3 Afghanistan AFG   1963                                 0.258      986.
9.54e6
## 4 Afghanistan AFG   1964                                 0.256     1083.
9.74e6
## 5 Afghanistan AFG   1965                                 0.385     1099.
9.96e6
## 6 Afghanistan AFG   1966                                 0.511     1012.
1.02e7
## # … with abbreviated variable names
## #   ¹`Cereal yield (kilograms per hectare) (kg per hectare)`,
## #   ²`Total population (Gapminder)`

head(land_use)

## # A tibble: 6 × 6
##   Entity      Code  Year  `Cereal yield index` Change to land area use…¹
Total…²
```

```
##   <chr>       <chr> <chr>                <dbl>                      <dbl>
<dbl>
## 1 Afghanistan AFG   1961                   100                        100
9.17e6
## 2 Afghanistan AFG   1962                    97                        103
9.35e6
## 3 Afghanistan AFG   1963                    88                        103
9.54e6
## 4 Afghanistan AFG   1964                    97                        104
9.74e6
## 5 Afghanistan AFG   1965                    99                        104
9.96e6
## 6 Afghanistan AFG   1966                    91                        104
1.02e7
## # … with abbreviated variable names
## #   ¹`Change to land area used for cereal production since 1961`,
## #   ²`Total population (Gapminder)`
```

Lets check the column names, as I will need to use these in visualization

```
names(key_crop_yields)
```

```
##  [1] "Entity"                           "Code"
##  [3] "Year"                             "Wheat (tonnes per hectare)"
##  [5] "Rice (tonnes per hectare)"        "Maize (tonnes per hectare)"
##  [7] "Soybeans (tonnes per hectare)"    "Potatoes (tonnes per hectare)"
##  [9] "Beans (tonnes per hectare)"       "Peas (tonnes per hectare)"
## [11] "Cassava (tonnes per hectare)"     "Barley (tonnes per hectare)"
## [13] "Cocoa beans (tonnes per hectare)" "Bananas (tonnes per hectare)"
```

```
names(tractors)
```

```
## [1] "Entity"
## [2] "Code"
## [3] "Year"
## [4] "Tractors per 100 sq km arable land"
## [5] "Cereal yield (kilograms per hectare) (kg per hectare)"
## [6] "Total population (Gapminder)"
```

```
names(land_use)
```

```
## [1] "Entity"
## [2] "Code"
## [3] "Year"
## [4] "Cereal yield index"
## [5] "Change to land area used for cereal production since 1961"
## [6] "Total population (Gapminder)"
```

Rename Columns

```
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)`)
```

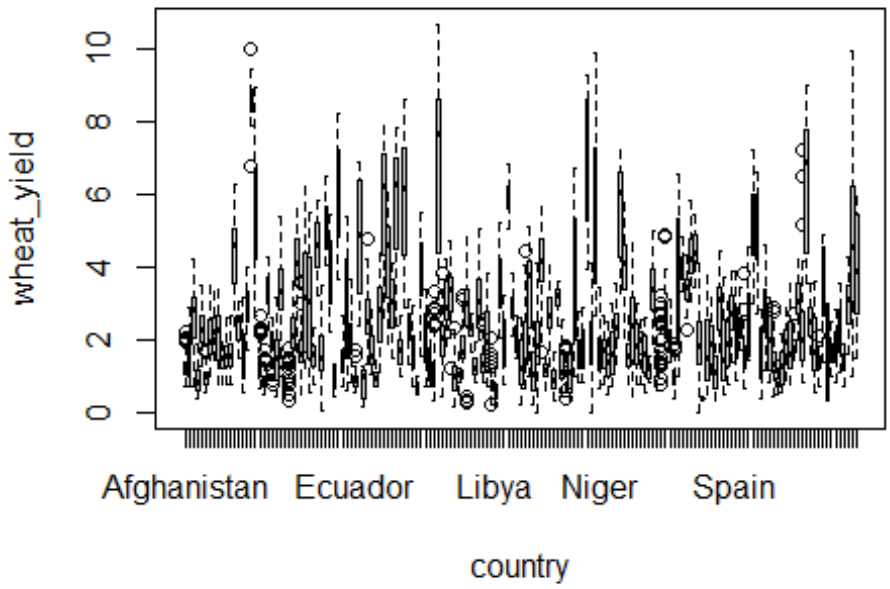
Now rename the Dataset 2: cereal_yields_vs_tractor_inputs_in_agriculture.csv

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

names(land_use) <- c("Entity", "Code", "Year", "Cereal_yield_index",
"Change_in_land_area_cereal", "Total_population")

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

Statistics of the DataSets

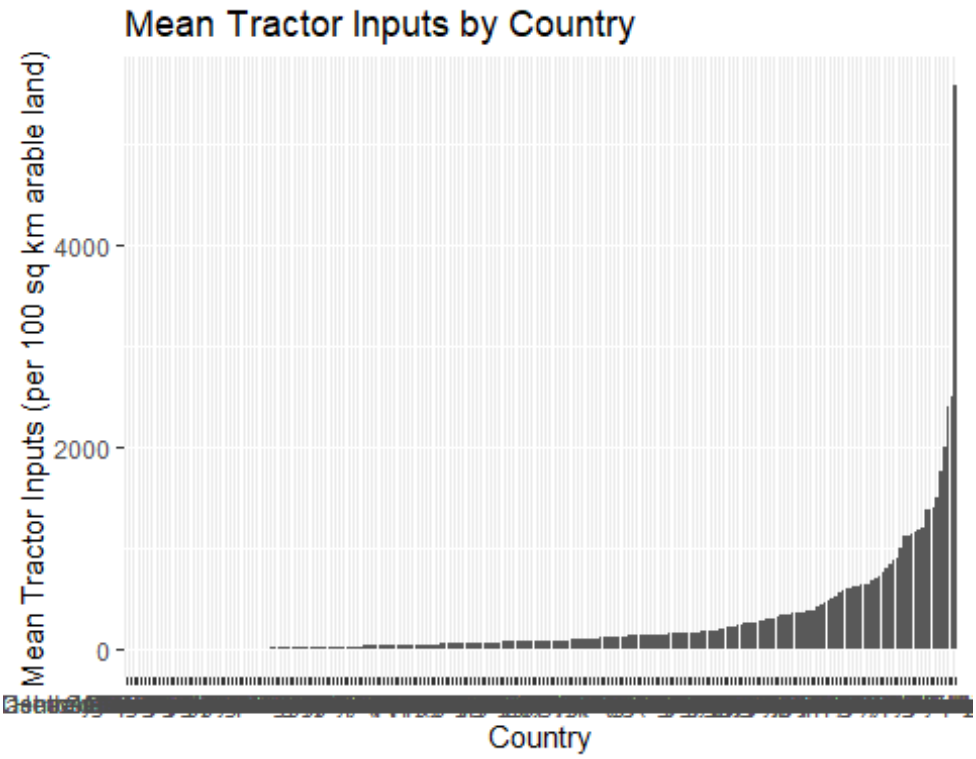
For First Dataset

```
# 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

```
# 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)

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

```
# 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
```

```
##                                                 Entity Cereal_yield_index
## 1                                          Afghanistan          120.59259
## 2                                              Albania          319.55556
## 3                                              Algeria          256.88889
## 4                                               Angola           73.83333
## 5                                           Arab World          168.61111
## 6                                            Argentina          192.12963
## 7                                            Australia          142.64815
## 8                                              Austria          193.24074
## 9                                              Bahamas          345.11111
```

```
## 10                                        Bangladesh          154.00000
## 11                                          Barbados          135.92593
## 12                                            Belize          330.38889
## 13                                             Benin          160.20370
## 14                                            Bhutan          114.18519
## 15                                           Bolivia          147.90741
## 16                                          Botswana          104.92593
## 17                                            Brazil          165.94444
## 18                                            Brunei           82.00000
## 19                                          Bulgaria          188.37037
## 20                                      Burkina Faso          180.35185
## 21                                           Burundi          125.11111
## 22                                          Cambodia          148.98148
## 23                                          Cameroon          137.25926
## 24                                            Canada          249.72222
## 25                                        Cape Verde           59.29630
## 26                           Caribbean small states          159.01852
## 27                         Central African Republic          198.72222
## 28                 Central Europe and the Baltics          178.98148
## 29                                              Chad          108.51852
## 30                                             Chile          247.31481
## 31                                             China          314.75926
## 32                                          Colombia          202.37037
## 33                                           Comoros          108.75926
## 34                                             Congo          110.24074
## 35                                        Costa Rica          228.87037
## 36                                      Cote d'Ivoire          192.81481
## 37                                              Cuba          179.94444
## 38                                            Cyprus          245.79630
## 39                 Democratic Republic of Congo          109.51852
## 40                                           Denmark          147.70370
## 41                                          Dominica          114.14815
## 42                                Dominican Republic          191.53704
## 43                  Early-demographic dividend          189.37037
## 44                           East Asia & Pacific          230.07407
## 45  East Asia & Pacific (excluding high income)          261.87037
## 46               East Asia & Pacific (IDA & IBRD)          264.74074
## 47                                          Ecuador          181.68519
## 48                                            Egypt          183.33333
## 49                                      El Salvador          200.59259
## 50                                         Ethiopia          162.20370
## 51                                         Euro area          213.27778
## 52                          Europe & Central Asia          166.61111
## 53 Europe & Central Asia (excluding high income)          169.35185
## 54             Europe & Central Asia (IDA & IBRD)          157.22222
## 55                                  European Union          196.64815
## 56                                            Fiji          121.61111
## 57                                         Finland          150.24074
## 58     Fragile and conflict affected situations          130.00000
## 59                                          France          239.83333
```

```
## 60                                                 Gabon           97.66667
## 61                                                Gambia          100.96296
## 62                                               Germany          211.61111
## 63                                                 Ghana          136.09259
## 64                                                Greece          249.57407
## 65                                               Grenada          109.24074
## 66                                                  Guam          111.42593
## 67                                             Guatemala          193.27778
## 68                                                Guinea          106.03704
## 69                                         Guinea-Bissau          134.51852
## 70                                                Guyana          157.61111
## 71                                                 Haiti           97.74074
## 72           Heavily indebted poor countries (HIPC)          120.14815
## 73                                           High income          181.01852
## 74                                              Honduras          127.35185
## 75                                             Hong Kong           85.62963
## 76                                               Hungary          220.27778
## 77                                             IBRD only          210.72222
## 78                                         IDA & IBRD total          196.37037
## 79                                             IDA blend          174.61111
## 80                                              IDA only          133.25926
## 81                                             IDA total          143.51852
## 82                                                 India          188.01852
## 83                                             Indonesia          214.40741
## 84                                                  Iran          173.59259
## 85                                                  Iraq          115.35185
## 86                                               Ireland          179.92593
## 87                                                Israel          216.24074
## 88                                                 Italy          182.68519
## 89                                               Jamaica          157.09259
## 90                                                 Japan          133.33333
## 91                                                Jordan          179.40741
## 92                                                 Kenya          120.24074
## 93                                                  Laos          252.96296
## 94                          Late-demographic dividend          221.24074
## 95                          Latin America & Caribbean          183.24074
## 96   Latin America & Caribbean (excluding high income)          181.22222
## 97            Latin America & Caribbean (IDA & IBRD)          183.31481
## 98       Least developed countries: UN classification          130.16667
## 99                                              Lebanon          162.18519
## 100                                             Lesotho           91.74074
## 101                                             Liberia          197.33333
## 102                                               Libya          253.22222
## 103                           Low & middle income          197.50000
## 104                                          Low income          120.90741
## 105                           Lower middle income          183.18519
## 106                                          Madagascar          115.74074
## 107                                              Malawi          127.55556
## 108                                            Malaysia          135.88889
## 109                                             Maldives          140.18519
```

```
## 110                                          Mali          134.94444
## 111                                          Malta          232.27778
## 112                                       Mauritania          176.12963
## 113                                        Mauritius          262.51852
## 114                                           Mexico          210.29630
## 115                            Middle East & North Africa          196.31481
## 116 Middle East & North Africa (excluding high income)          194.12963
## 117            Middle East & North Africa (IDA & IBRD)          194.12963
## 118                                     Middle income          205.00000
## 119                                          Mongolia          228.44444
## 120                                           Morocco          256.64815
## 121                                        Mozambique           83.09259
## 122                                           Myanmar          170.79630
## 123                                           Namibia           94.96296
## 124                                             Nepal          104.27778
## 125                                       Netherlands          172.37037
## 126                                     New Caledonia          206.42593
## 127                                       New Zealand          163.77778
## 128                                         Nicaragua          161.59259
## 129                                             Niger           81.27778
## 130                                           Nigeria          150.98148
## 131                                     North America          191.38889
## 132                                       North Korea          135.50000
## 133                                            Norway          125.00000
## 134                                      OECD members          181.57407
## 135                                              Oman          240.14815
## 136                                Other small states          138.92593
## 137                           Pacific island small states          118.33333
## 138                                          Pakistan          209.79630
## 139                                            Panama          171.29630
## 140                                  Papua New Guinea          100.37037
## 141                                          Paraguay          150.09259
## 142                                              Peru          170.25926
## 143                                       Philippines          205.53704
## 144                                            Poland          154.14815
## 145                                          Portugal          250.83333
## 146                            Post-demographic dividend          177.92593
## 147                             Pre-demographic dividend          120.29630
## 148                                       Puerto Rico          259.87037
## 149                                           Romania          179.50000
## 150                                            Rwanda          142.14815
## 151                                       Saint Lucia           27.38889
## 152                       Saint Vincent and the Grenadines          184.00000
## 153                                  Sao Tome and Principe          121.68519
## 154                                      Saudi Arabia          223.20370
## 155                                           Senegal          140.22222
## 156                                      Sierra Leone          139.42593
## 157                                      Small states          144.75926
## 158                                    Solomon Islands          164.62963
## 159                                           Somalia          117.35185
```

```
## 160                                        South Africa          193.77778
## 161                                          South Asia          181.53704
## 162                               South Asia (IDA & IBRD)          181.53704
## 163                                         South Korea          165.42593
## 164                                               Spain          220.24074
## 165                                           Sri Lanka          157.46296
## 166                                  Sub-Saharan Africa          130.29630
## 167          Sub-Saharan Africa (excluding high income)          130.29630
## 168                      Sub-Saharan Africa (IDA & IBRD)          130.29630
## 169                                               Sudan           66.98148
## 170                                            Suriname          136.11111
## 171                                           Swaziland          260.85185
## 172                                              Sweden          142.75926
## 173                                         Switzerland          185.25926
## 174                                               Syria          217.51852
## 175                                            Tanzania          145.20370
## 176                                            Thailand          138.01852
## 177                                               Timor          114.18519
## 178                                                Togo          187.03704
## 179                                 Trinidad and Tobago          105.85185
## 180                                             Tunisia          212.85185
## 181                                              Turkey          200.05556
## 182                                              Uganda          157.29630
## 183                                      United Kingdom          175.03704
## 184                                       United States          187.31481
## 185                                 Upper middle income          223.27778
## 186                                             Uruguay          281.62963
## 187                                             Vanuatu          103.61111
## 188                                           Venezuela          211.53704
## 189                                             Vietnam          170.79630
## 190                                               World          184.16667
## 191                                               Yemen          114.51852
## 192                                              Zambia          186.64815
## 193                                            Zimbabwe          116.72222
```

8- Data Visualization

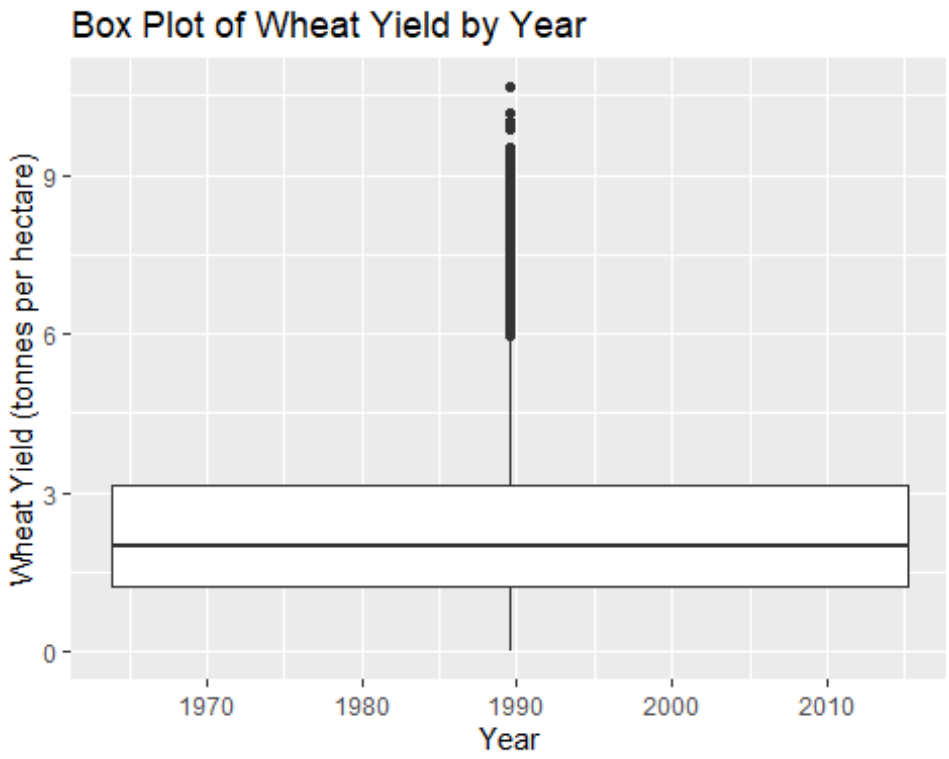
1. Line chart of yield trends over time for select crops and countries:

```
names(key_crop_yields)
```

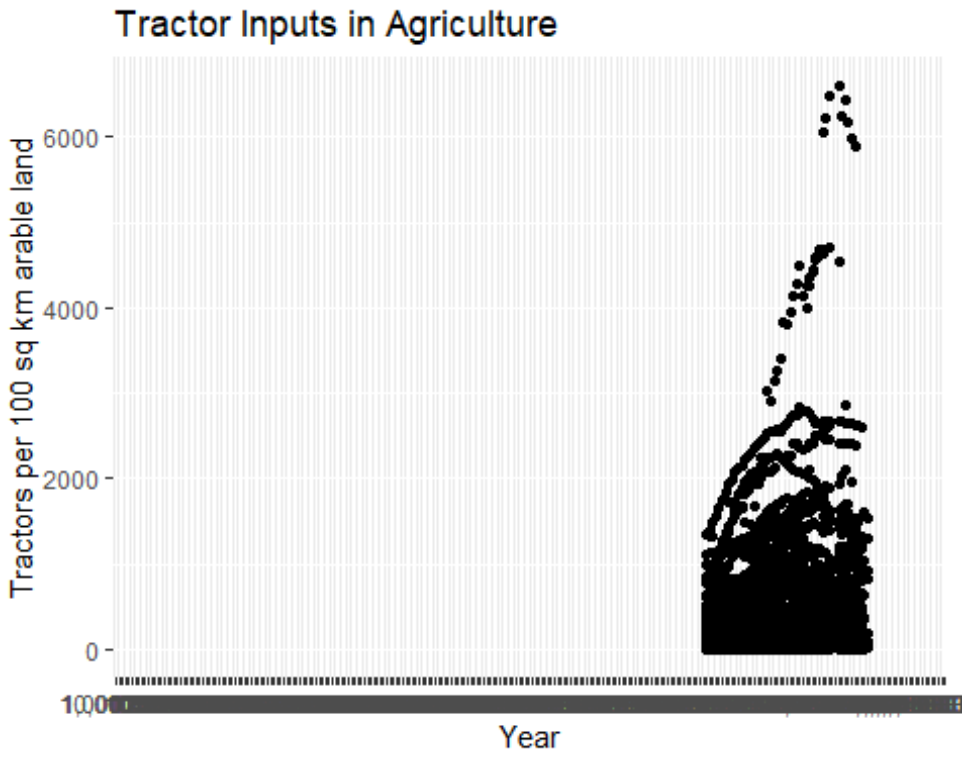
```
##  [1] "country"       "code"          "year"          "wheat_yield"
##  [5] "rice_yield"    "maize_yield"   "soybean_yield" "potato_yield"
##  [9] "beans_yield"   "peas_yield"    "cassava_yield" "barley_yield"
## [13] "cocoa_yield"   "banana_yield"
```

```
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")
```

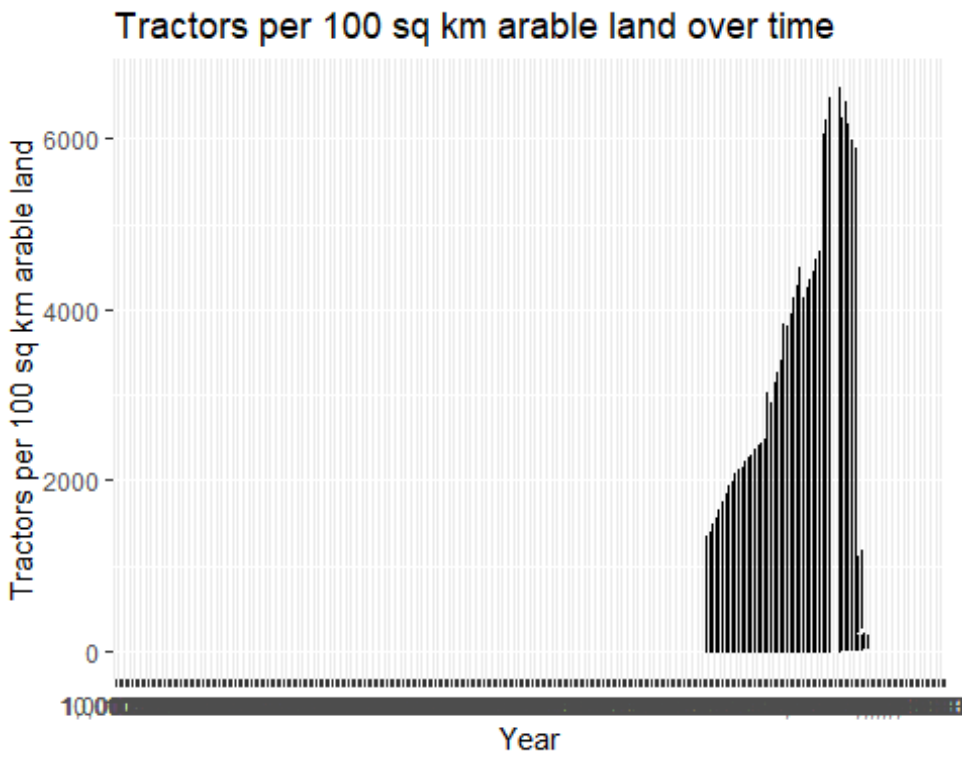
```
## Warning: Continuous x aesthetic
## ℹ did you forget `aes(group = ...)`?

## Warning: Removed 4974 rows containing non-finite values
(`stat_boxplot()`).
```



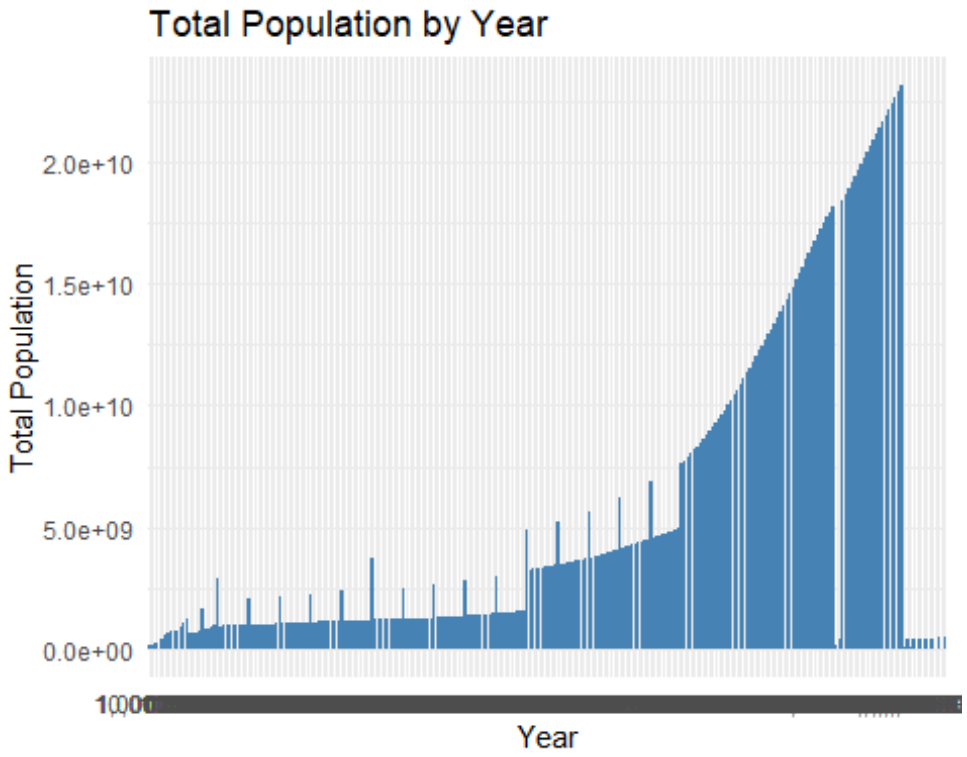

```
ggplot(tractors, aes(x=Year, y=Tractor_per_hundred)) +
  geom_point() +
  labs(title = "Tractor Inputs in Agriculture", x = "Year", y = "Tractors per
100 sq km arable land")

## Warning: Removed 41911 rows containing missing values (`geom_point()`).
```




```
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")

## Warning: Removed 35472 rows containing missing values (`geom_line()`).
```

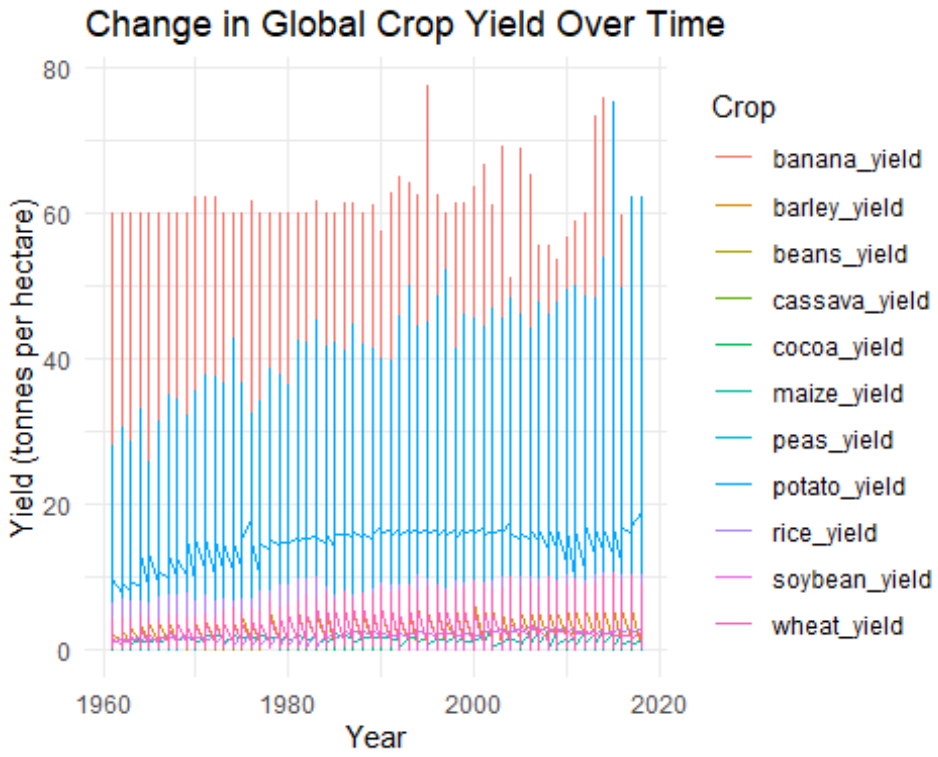



```
ggplot(data = land_use, aes(x = Year, y = Total_population)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Total Population by Year", x = "Year", y = "Total
Population") +
  theme_minimal()

## Warning: Removed 2376 rows containing missing values (`position_stack()`).
```



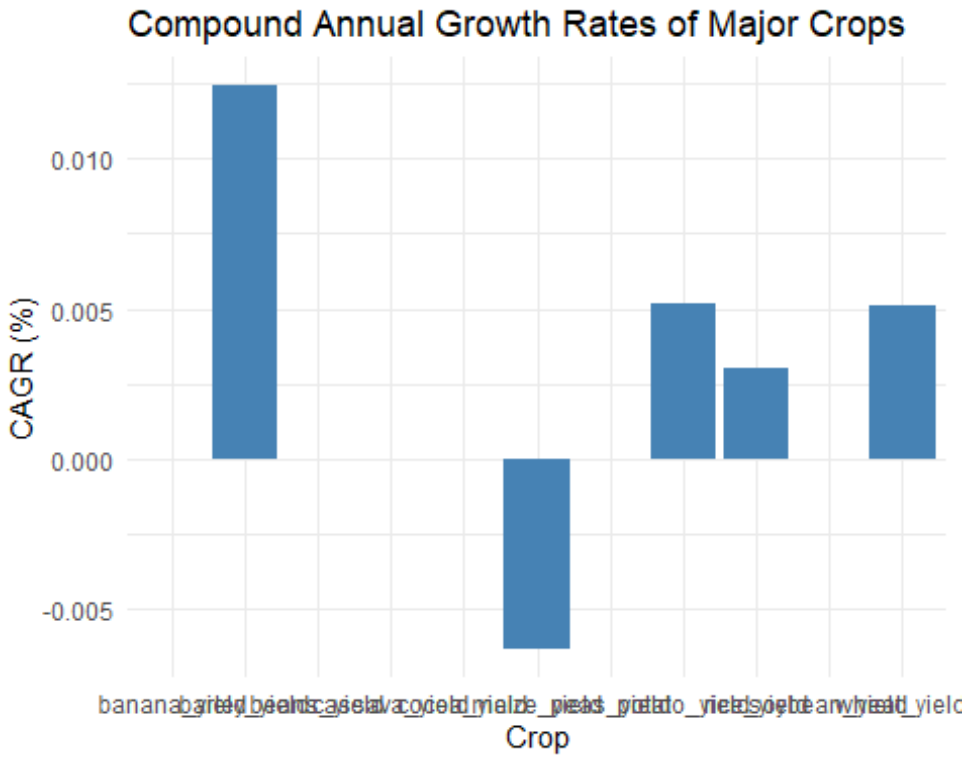
9- Visualizations for Research Questions *Research Question 1*:

```
# 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")

# 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()

## Warning: Removed 11 rows containing missing values (`geom_line()`).
```

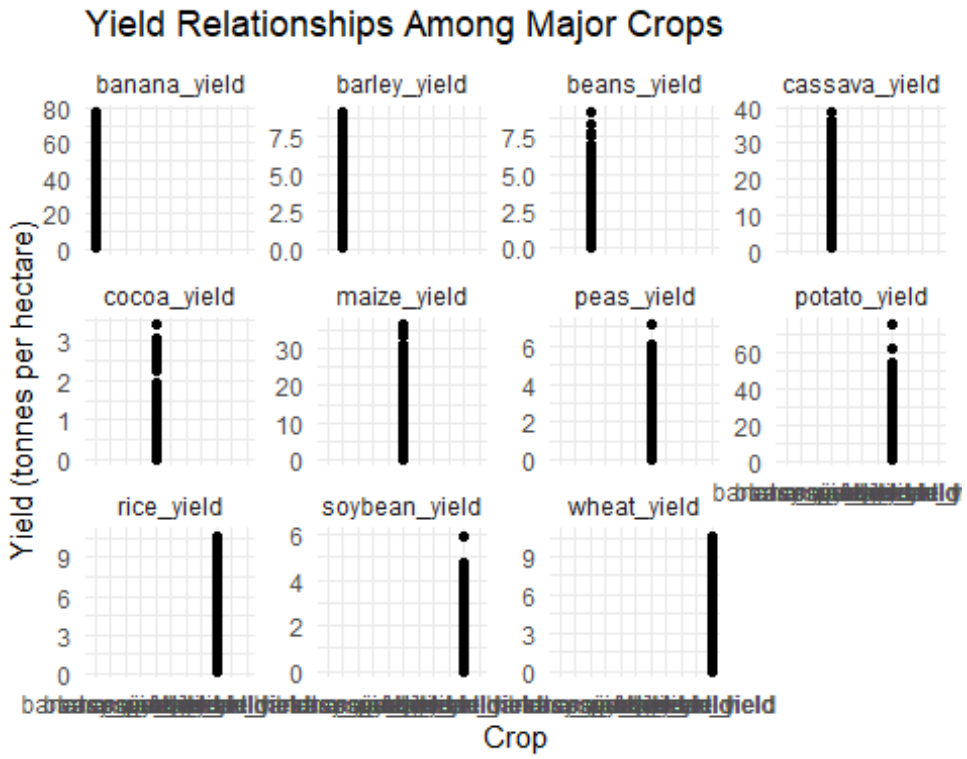



```r
# calculate the compound annual growth rate (CAGR) for each crop
crop_growth_rates <- crop_yields_long %>%
  group_by(Crop) %>%
  summarize(CAGR = ((last(Yield)/first(Yield))^(1/n()) - 1) * 100)

# 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()
```

```
## Warning: Removed 6 rows containing missing values (`position_stack()`).
```




```
# 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")

## `geom_smooth()` using formula = 'y ~ x'

## Warning: Removed 58819 rows containing non-finite values
(`stat_smooth()`).

## Warning: Removed 58819 rows containing missing values (`geom_point()`).
```



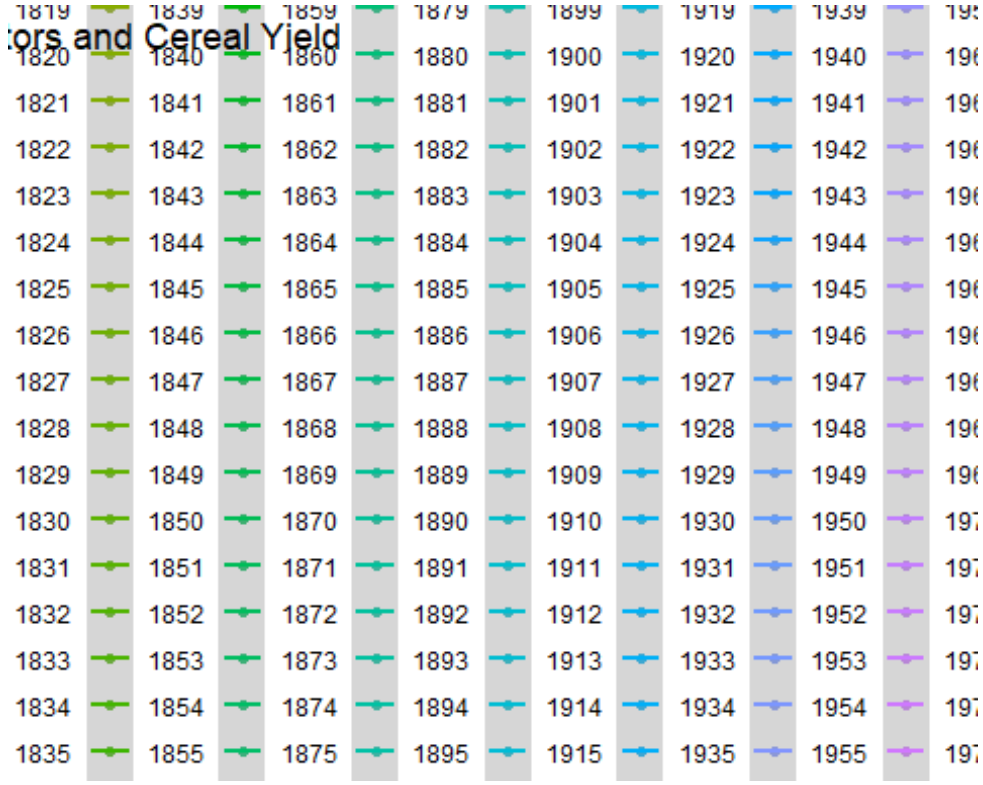
Research Question 2:

```
# 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()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 42501 rows containing non-finite values
(`stat_smooth()`).
```

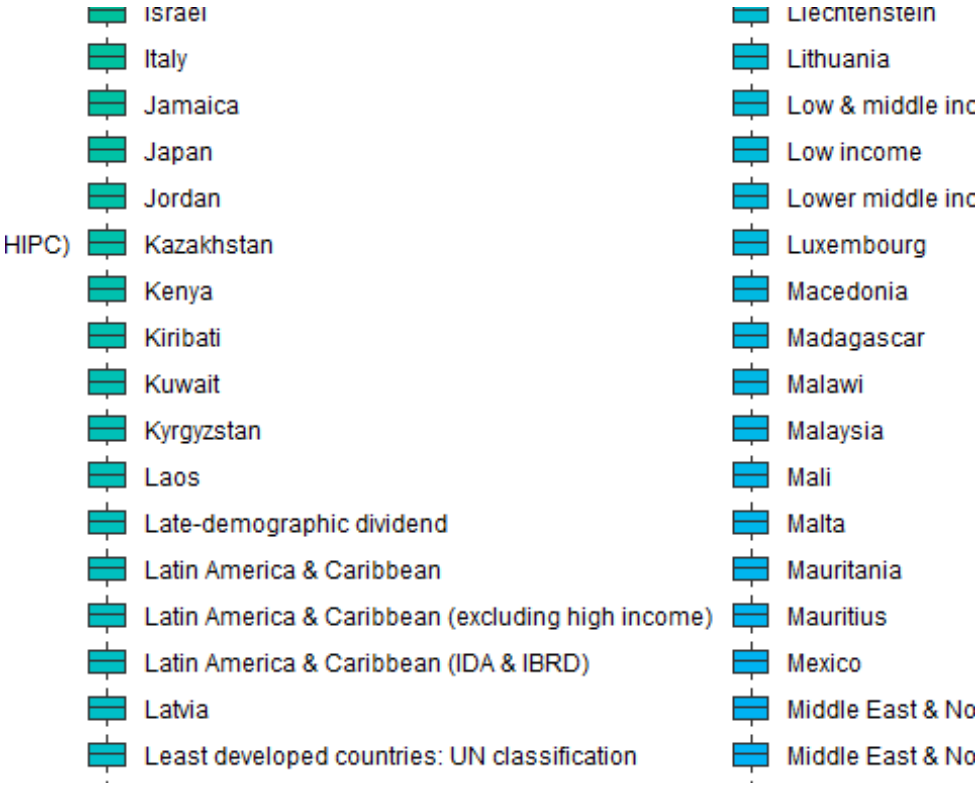
```
## Warning: Removed 42501 rows containing missing values (`geom_point()`).
```




```
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()
```

```
## Warning: Removed 37763 rows containing non-finite values
(`stat_boxplot()`).
```

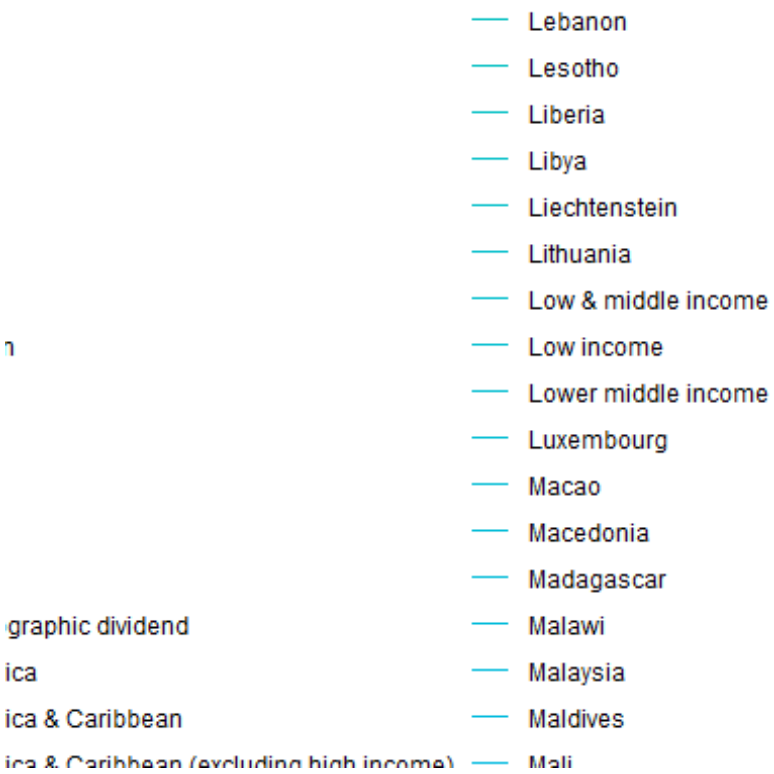

```
# Plot box plots of tractors and cereal yield
ggplot(tractors, aes(x = Year, y = `Tractor_per_hundred`, fill = Entity)) +
  geom_boxplot() +
  labs(title = "Distribution of Tractors per 100 sq km arable land over
time",
       x = "Year",
       y = "Tractors per 100 sq km arable land",
       fill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

## Warning: Removed 41911 rows containing non-finite values
(`stat_boxplot()`).
```



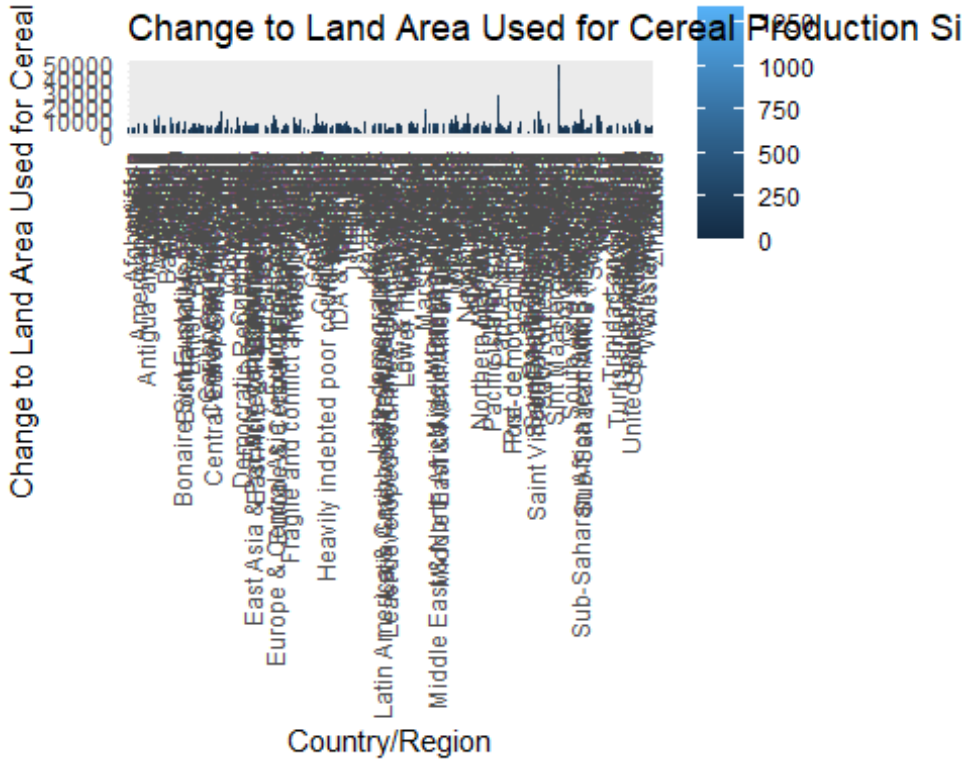
Research Question 3

```
# 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()

## `geom_smooth()` using formula = 'y ~ x'

## Warning: Removed 38837 rows containing non-finite values
(`stat_smooth()`).

## Warning: Removed 38837 rows containing missing values (`geom_point()`).
```


```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()
```




```
# 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))

## Warning: Removed 38837 rows containing missing values
(`position_stack()`).
```



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