GRAIN

Syed Ali

2023-03-08

#install.packages("readxl")  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.4.0 ✔ purrr 1.0.1  
## ✔ tibble 3.1.8 ✔ dplyr 1.1.0  
## ✔ tidyr 1.3.0 ✔ 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(readxl)  
data <- read\_excel("C:/Users/SOFTWARE ENGINEER/Desktop/syed ali/New/GRAIN---Land-grab-deals---Jan-2012-2.xlsx")  
data

## # A tibble: 159 × 10  
## Landgrab…¹ Landg…² Base Sector Hecta…³ Produ…⁴ Proje…⁵ Year Statu…⁶ Summary  
## <chr> <chr> <chr> <chr> <dbl> <chr> <chr> <dbl> <chr> <chr>   
## 1 Argentina Adecoa… US Agrib… 242000 Cattle… <NA> 0 Done Adecoa…  
## 2 Uruguay Adecoa… US Agrib… 8600 Cattle… <NA> 0 Done Adecoa…  
## 3 Algeria Al Qud… UAE real … 31000 Milk, … <NA> 17 Done Al Qud…  
## 4 New Zeala… Ingleb… Denm… Finan… 14461 Cattle… <NA> 1901 Done The In…  
## 5 Australia Ho Myo… Sout… Indus… 216000 Cattle… <NA> 1999 Done Ho Myo…  
## 6 Australia JBS Braz… AB 1876 Livest… US$40 … 1999 Done In 200…  
## 7 Australia Terra … UK Finan… 3200000 Livest… US$456… 1999 Done In 200…  
## 8 Australia Hassad… Qatar Agrib… 750000 Sheep,… US$400… 1999 Don Hassad…  
## 9 Australia Wilmar… Sing… AB 2500 Sugar … US$1,8… 1999 Done Wilmar…  
## 10 Brazil Aquila Germ… Finan… 250000 Cattle… <NA> 2000 In pro… Aquila…  
## # … with 149 more rows, and abbreviated variable names ¹​Landgrabbed,  
## # ²​Landgrabber, ³​Hectares, ⁴​Production, ⁵​`Projected investment`,  
## # ⁶​`Status of deal`

Before we start cleaning the data, let’s take a quick look at the data structure and see what needs to be cleaned.

We can use the following commands to get a quick overview of the data:

head(data)

## # A tibble: 6 × 10  
## Landgrabbed Landg…¹ Base Sector Hecta…² Produ…³ Proje…⁴ Year Statu…⁵ Summary  
## <chr> <chr> <chr> <chr> <dbl> <chr> <chr> <dbl> <chr> <chr>   
## 1 Argentina Adecoa… US Agrib… 242000 Cattle… <NA> 0 Done Adecoa…  
## 2 Uruguay Adecoa… US Agrib… 8600 Cattle… <NA> 0 Done Adecoa…  
## 3 Algeria Al Qud… UAE real … 31000 Milk, … <NA> 17 Done Al Qud…  
## 4 New Zealand Ingleb… Denm… Finan… 14461 Cattle… <NA> 1901 Done The In…  
## 5 Australia Ho Myo… Sout… Indus… 216000 Cattle… <NA> 1999 Done Ho Myo…  
## 6 Australia JBS Braz… AB 1876 Livest… US$40 … 1999 Done In 200…  
## # … with abbreviated variable names ¹​Landgrabber, ²​Hectares, ³​Production,  
## # ⁴​`Projected investment`, ⁵​`Status of deal`

str(data)

## tibble [159 × 10] (S3: tbl\_df/tbl/data.frame)  
## $ Landgrabbed : chr [1:159] "Argentina" "Uruguay" "Algeria" "New Zealand" ...  
## $ Landgrabber : chr [1:159] "Adecoagro" "Adecoagro" "Al Qudra" "Ingleby Company" ...  
## $ Base : chr [1:159] "US" "US" "UAE" "Denmark" ...  
## $ Sector : chr [1:159] "Agribusiness" "Agribusiness" "real estate, Finance" "Finance" ...  
## $ Hectares : num [1:159] 242000 8600 31000 14461 216000 ...  
## $ Production : chr [1:159] "Cattle, dairy, grains, soybeans" "Cattle, grains, soybeans" "Milk, olive oil, potatoes" "Cattle, sheep" ...  
## $ Projected investment: chr [1:159] NA NA NA NA ...  
## $ Year : num [1:159] 0 0 17 1901 1999 ...  
## $ Status of deal : chr [1:159] "Done" "Done" "Done" "Done" ...  
## $ Summary : chr [1:159] "Adecoagro is a publicly traded company launched and controlled by the fund of US billionaire George Soros, with"| \_\_truncated\_\_ "Adecoagro is a publicly traded company launched and controlled by the fund of US billionaire George Soros, with"| \_\_truncated\_\_ "Al Qudra Holding is a joint-stock company established in Abu Dhabi in 2005. In February 2008, the company told "| \_\_truncated\_\_ "The Ingleby Company, which is owned by Denmark's Rausing dynasty, began purchasing farms in New Zealand in 1999"| \_\_truncated\_\_ ...

summary(data)

## Landgrabbed Landgrabber Base Sector   
## Length:159 Length:159 Length:159 Length:159   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## Hectares Production Projected investment Year   
## Min. : 200 Length:159 Length:159 Min. : 0   
## 1st Qu.: 8481 Class :character Class :character 1st Qu.:2012   
## Median : 20110 Mode :character Mode :character Median :2013   
## Mean : 97127 Mean :1974   
## 3rd Qu.: 92500 3rd Qu.:2014   
## Max. :3200000 Max. :2014   
## NA's :2   
## Status of deal Summary   
## Length:159 Length:159   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##   
##

## Based on the output of head(), str(), and summary() functions, here are some observations and potential issues in the data:

The Hectares variable is of numerical data type, which is good. However, there seems to be a very wide range of values, with the minimum at 200 hectares and the maximum at 3,200,000 hectares. This suggests that there may be outliers in the data that need to be investigated and potentially removed.

The Year variable is also of numerical data type, which is good. However, the minimum value is 0, which is not a valid year. This suggests that there may be missing or incorrect data in this variable.

The Projected investment variable has a lot of missing values (indicated by “NA” in the output). Depending on the analysis, these missing values may need to be handled or imputed in some way.

The Landgrabbed, Landgrabber, Base, Sector, Production, Status of deal, and Summary variables are all of character data type, which is reasonable given their nature as textual descriptions. However, there may be formatting or spelling errors in these variables that need to be checked and potentially corrected.

The Sector variable seems to have multiple values separated by commas in some cases. This may cause issues in data analysis and may need to be addressed, such as by splitting the values into separate variables or categories.

Based on these observations, some potential data cleaning tasks that could be performed include:

Investigating and potentially removing outliers in the Hectares variable.

Checking and correcting any missing or incorrect data in the Year variable.

Handling or imputing missing values in the Projected investment variable, depending on the analysis.

Checking and correcting formatting or spelling errors in the Landgrabbed, Landgrabber, Base, Sector, Production, Status of deal, and Summary variables.

Addressing any issues with the Sector variable that may affect data analysis.

## Clean the data

### Convert year to numeric

data$Year <- as.numeric(data$Year)

### Remove rows with missing values

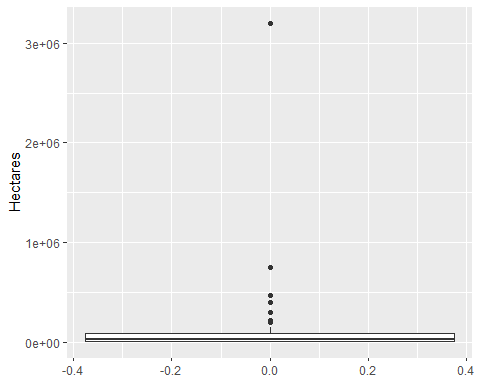
# count missing values  
sum(is.na(data))

## [1] 116

# remove rows with missing values  
data <- data %>%   
 drop\_na()

#### Outliers

# detect outliers using boxplot  
ggplot(data, aes(y = Hectares)) +   
 geom\_boxplot()



# remove outliers  
data <- data %>%   
 filter(Hectares <= quantile(Hectares, 0.99))

### Spelling and formatting errors

# fix spelling errors in Landgrabber column  
data$Landgrabber <- ifelse(data$Landgrabber == "Adecoagro ", "Adecoagro", data$Landgrabber)

## Save the cleaned data

write.csv(data, "C:/Users/SOFTWARE ENGINEER/Desktop/syed ali/New/data.csv", row.names = FALSE)

head(data)

## # A tibble: 6 × 10  
## Landgrabbed Landg…¹ Base Sector Hecta…² Produ…³ Proje…⁴ Year Statu…⁵ Summary  
## <chr> <chr> <chr> <chr> <dbl> <chr> <chr> <dbl> <chr> <chr>   
## 1 Australia JBS Braz… AB 1876 Livest… US$40 … 1999 Done In 200…  
## 2 Australia Hassad… Qatar Agrib… 750000 Sheep,… US$400… 1999 Don Hassad…  
## 3 Australia Wilmar… Sing… AB 2500 Sugar … US$1,8… 1999 Done Wilmar…  
## 4 Colombia China China Gover… 400000 Cereals 0 2000 Propos… In 201…  
## 5 Ethiopia BHO Ag… India Agrib… 27000 Cereal… US$8/h… 2000 Done In May…  
## 6 Angola AfriAg… Port… Finan… 5000 Oil pa… US$30-… 2012 Done AfriAg…  
## # … with abbreviated variable names ¹​Landgrabber, ²​Hectares, ³​Production,  
## # ⁴​`Projected investment`, ⁵​`Status of deal`

str(data)

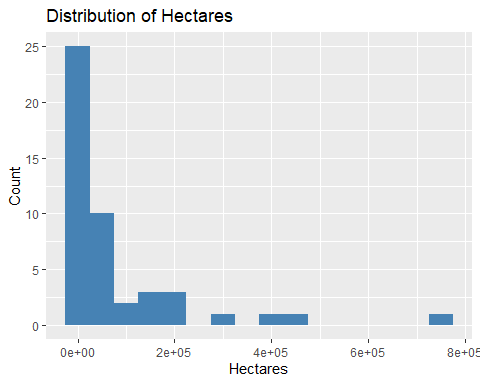
## tibble [47 × 10] (S3: tbl\_df/tbl/data.frame)  
## $ Landgrabbed : chr [1:47] "Australia" "Australia" "Australia" "Colombia" ...  
## $ Landgrabber : chr [1:47] "JBS" "Hassad Food" "Wilmar International" "China" ...  
## $ Base : chr [1:47] "Brazil" "Qatar" "Singapore" "China" ...  
## $ Sector : chr [1:47] "AB" "Agribusiness" "AB" "Government" ...  
## $ Hectares : num [1:47] 1876 750000 2500 400000 27000 ...  
## $ Production : chr [1:47] "Livestock" "Sheep, wheat" "Sugar cane" "Cereals" ...  
## $ Projected investment: chr [1:47] "US$40 million" "US$400 million" "US$1,876 million" "0" ...  
## $ Year : num [1:47] 1999 1999 1999 2000 2000 ...  
## $ Status of deal : chr [1:47] "Done" "Don" "Done" "Proposed" ...  
## $ Summary : chr [1:47] "In 2007, Brazilian meat giant JBS acquired US-based Swift Food & Co., giving it control over ten meat plants an"| \_\_truncated\_\_ "Hassad Food is a US$1-billion company established by Qatar's sovereign wealth fund. It has been leading the cou"| \_\_truncated\_\_ "Wilmar, one of the world's largest agricultural commodity producers and traders, bought Sucrogen Ltd, wich owns"| \_\_truncated\_\_ "In 2010, Portafolio reported that the Chinese Ambassador, accompanied by Chinese business representatives, made"| \_\_truncated\_\_ ...

summary(data)

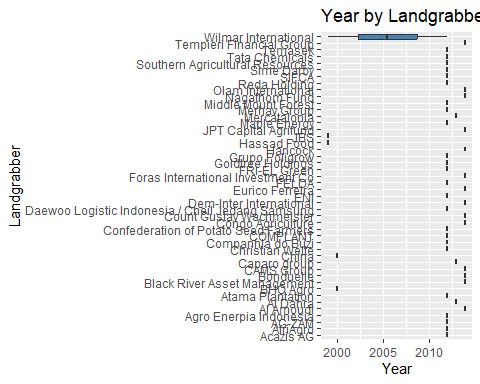
## Landgrabbed Landgrabber Base Sector   
## Length:47 Length:47 Length:47 Length:47   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Hectares Production Projected investment Year   
## Min. : 1876 Length:47 Length:47 Min. :1999   
## 1st Qu.: 8640 Class :character Class :character 1st Qu.:2012   
## Median : 24000 Mode :character Mode :character Median :2012   
## Mean : 81253 Mean :2011   
## 3rd Qu.: 75000 3rd Qu.:2014   
## Max. :750000 Max. :2014   
## Status of deal Summary   
## Length:47 Length:47   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##

## Visualizations

library(ggplot2)  
  
# histogram of Hectares  
ggplot(data, aes(x = Hectares)) +   
 geom\_histogram(binwidth = 50000, fill = "steelblue") +   
 labs(title = "Distribution of Hectares", x = "Hectares", y = "Count")



# boxplot of Year by Landgrabber  
ggplot(data, aes(x = Landgrabber, y = Year)) +   
 geom\_boxplot(fill = "steelblue") +   
 coord\_flip() +   
 labs(title = "Year by Landgrabber", x = "Landgrabber", y = "Year")



# scatterplot of Hectares vs. Projected investment  
ggplot(data, aes(x = Hectares, y = as.numeric(`Projected investment`))) +   
 geom\_point(color = "steelblue") +   
 scale\_y\_continuous(labels = scales::dollar\_format(prefix = "$")) +   
 labs(title = "Hectares vs. Projected Investment", x = "Hectares", y = "Projected Investment")

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning: Removed 45 rows containing missing values (`geom\_point()`).

