ANA 515 Assignment 3

Mounika Gangishetty

Reading the GRAIN data from downloaded excel

```
#  reading excel from local drive

setwd("/Users/vamshinaidi/Documents/ANA-515/PracticumProject")
excel_file  <- 'GRAINLandJan2012.xlsx'

# reading data from sheet1

sheet1 <- read_excel(excel_file,sheet="Sheet1")
head(sheet1)
```

```
## # A tibble: 6 x 10
##   Landgrabbed Landgrabber     Base        Sector             Hectares Production
##   <chr>       <chr>           <chr>       <chr>                 <dbl> <chr>
## 1 Argentina   Adecoagro       US          Agribusiness         242000 Cattle, d~
## 2 Uruguay     Adecoagro       US          Agribusiness           8600 Cattle, g~
## 3 Algeria     Al Qudra        UAE         real estate, Fina~    31000 Milk, oli~
## 4 New Zealand Ingleby Company Denmark     Finance               14461 Cattle, s~
## 5 Australia   Ho Myoung Farm  South Korea Industrial           216000 Cattle, g~
## 6 Australia   JBS             Brazil      AB                    1876 Livestock
## # i 4 more variables: `Projected investment` <chr>, Year <dbl>,
## #   `Status of deal` <chr>, Summary <chr>
```

```
# reading data from sheet2

sheet2 <- read_excel(excel_file,sheet="Sheet2")
head(sheet2)
```

```
## # A tibble: 6 x 10
##   Landgrabbed Landgrabber              Base      Sector      Hectares Production
##   <chr>       <chr>                    <chr>     <chr>          <dbl> <chr>
## 1 Angola      ENI                      NA        Energy         12000 Oil palm
## 2 Angola      Quifel Natural Resources Portugal  Agribusine~    10000 Oilseed
## 3 Argentina   Terra Magna Capital      Fran      Finance        70500 Crops
## 4 Argentina   DWS GALOF                Germany   Finance        20000 Crops
## 5 Argentina   Calyx Agro               France    Finance         5719 Crops (ma~
## 6 Arg         Siva Group               Singapore Agribusine~     2000 Olives
## # i 4 more variables: `Projected investment` <chr>, Year <chr>,
## #   `Status of deal` <chr>, Summary <chr>
```

Using unique() function to find the distinct values in the "Land-grabbed" column, i.e distinct country names from sheet1.

This data is about sales of vast amounts of agricultural land in less developed countries. The dataset has information about 52 countries agricultural land sales data. The datset has few null values for columns like 'Projected investment' ,'Year' and invalid data for columns like Status of Deal' and 'Year'. We are going to remove these outliers and anamolies from the dataset.

Removing missing values, null, NA and invalid data from the datset

```
# Removing the sales data in the sheets where projected investment amount is null,
# Hectares is null and  invalid years

filter_sheet1 <- sheet1 %>%
  filter(!is.na(`Projected investment`) & !is.null(Year) & !is.na(Year) & !is.na(`Hectares`))

head(filter_sheet1)
```

```
## # A tibble: 6 x 10
##   Landgrabbed Landgrabber          Base      Sector       Hectares Production
##   <chr>       <chr>                <chr>     <chr>           <dbl> <chr>
## 1 Australia   JBS                  Brazil    AB               1876 Livestock
## 2 Australia   Terra Firma Capital  UK        Finance       3200000 Livestock
## 3 Australia   Hassad Food          Qatar     Agribusiness   750000 Sheep, wheat
## 4 Australia   Wilmar International Singapore AB               2500 Sugar cane
## 5 Colombia    China                China     Government     400000 Cereals
## 6 Ethiopia    BHO Agro             India     Agribusiness    27000 Cereal, oils~
## # i 4 more variables: `Projected investment` <chr>, Year <dbl>,
## #   `Status of deal` <chr>, Summary <chr>
```

```
filter_sheet2 <- sheet2 %>%
  filter(!is.na(`Projected investment`) & !is.null(Year) & !is.na(Year) & !is.na(`Hectares`))


head(filter_sheet2)
```

```
## # A tibble: 6 x 10
##   Landgrabbed Landgrabber                Base        Sector  Hectares Production
##   <chr>       <chr>                      <chr>       <chr>      <dbl> <chr>
## 1 Angola      "Quifel Natural Resources" Portugal    Agribu~    10000 Oilseed
## 2 Cameroon    "Herakles Capital\r\n"     US          Finance    73000 Oil palm
## 3 Peru        "Ecoamerica"               South Korea Agribu~    72000 Crops, fo~
## 4 Russia      "Olam International"       Singapore   Agribu~    60000 Crops, da~
## 5 Angola      "CAMC Engineering Co. Ltd" China       Constr~     1500 Rice
## 6 Argentina   "Beidahuang"               China       Agribu~   320000 Maize, so~
## # i 4 more variables: `Projected investment` <chr>, Year <chr>,
## #   `Status of deal` <chr>, Summary <chr>
```

Correcting the misspelled data for data consistency

```
#
# Spell correction in column "Status of deal"
#
```

```r
# Printing the values before changes
#
head(filter_sheet1$'Status of deal')
```

```
## [1] "Done"     "Done"     "Don"      "Done"     "Proposed" "Done"
```

```r
head(filter_sheet2$'Status of deal')
```

```
## [1] "Don"        "Done"       "In process" "Done"       "Done"
## [6] "Suspended"
```

```r
#
# we are using mutate function to handle misspelled data ,replacing "Don" with correct spell "Done"
#  in the Status column and remaining we are not changing by passing the parameter TRUE in mutate funct


filter_sheet1 <- filter_sheet1 %>%
  mutate(`Status of deal` = case_when(
    `Status of deal` == "Don" ~ "Done",
    TRUE ~ `Status of deal`
  ))


filter_sheet2 <- filter_sheet2 %>%
  mutate(`Status of deal` = case_when(
    `Status of deal` == "Don" ~ "Done",
    TRUE ~ `Status of deal`
  ))

# Data after spell correction

head(filter_sheet1$'Status of deal')
```

```
## [1] "Done"     "Done"     "Done"     "Done"     "Proposed" "Done"
```

```r
head(filter_sheet2$'Status of deal')
```

```
## [1] "Done"       "Done"       "In process" "Done"       "Done"
## [6] "Suspended"
```

Combining the two data sheets into one

```r
# Combining the data from sheet1 and sheet2 into single file.

combined_data <- rbind(filter_sheet1,filter_sheet2)

head(combined_data)
```

```
## # A tibble: 6 x 10
##   Landgrabbed Landgrabber          Base      Sector       Hectares Production
##   <chr>       <chr>                <chr>     <chr>           <dbl> <chr>
## 1 Australia   JBS                  Brazil    AB               1876 Livestock
## 2 Australia   Terra Firma Capital  UK        Finance       3200000 Livestock
## 3 Australia   Hassad Food          Qatar     Agribusiness   750000 Sheep, wheat
## 4 Australia   Wilmar International Singapore AB               2500 Sugar cane
```

```
## 5 Colombia    China                China     Government     400000 Cereals
## 6 Ethiopia    BHO Agro             India     Agribusiness    27000 Cereal, oils~
## # i 4 more variables: `Projected investment` <chr>, Year <chr>,
## #   `Status of deal` <chr>, Summary <chr>
```

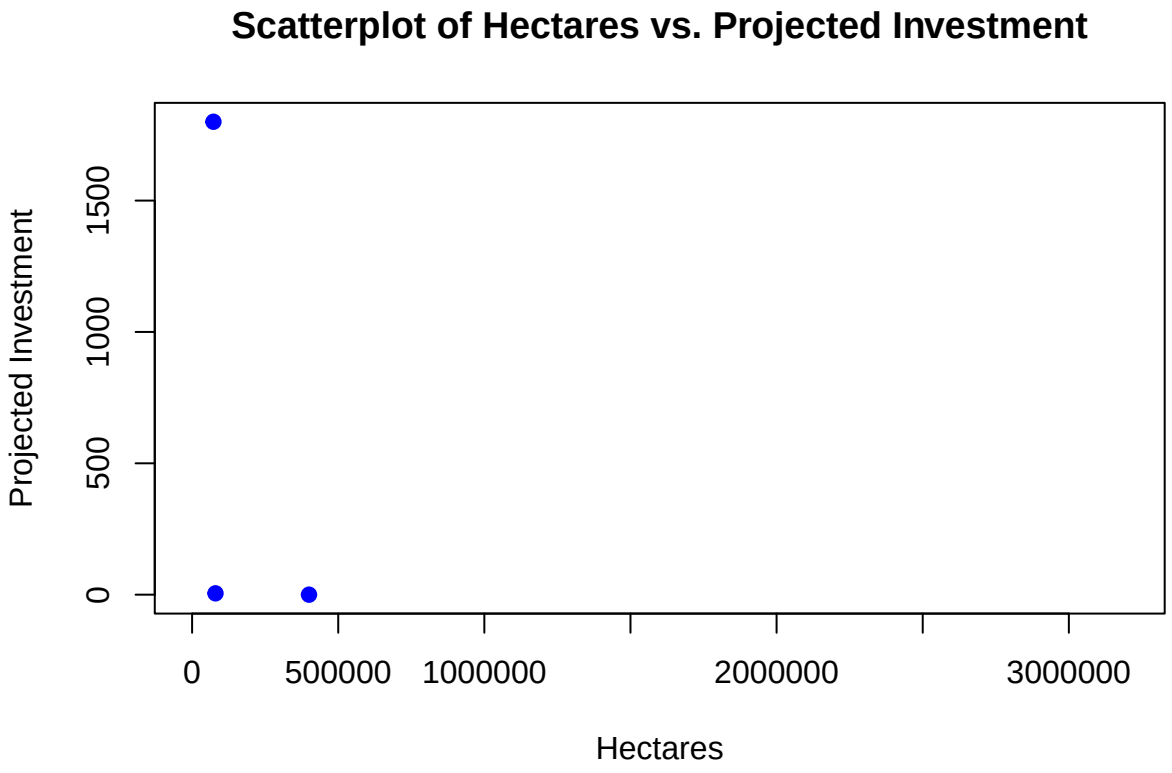
Sactter plot and histogram to understand hectares of land sales in less devloped countries among many years.

```
#
# scatterplot determining  "Hectares of Land" vs. "Projected investment"
# converting the values to numeric before proceeding to plot to avoid  "need finite 'xlim' values" erro
#

combined_data$Hectares <- as.numeric(combined_data$Hectares)
combined_data$`Projected investment` <- as.numeric(combined_data$`Projected investment`)
```

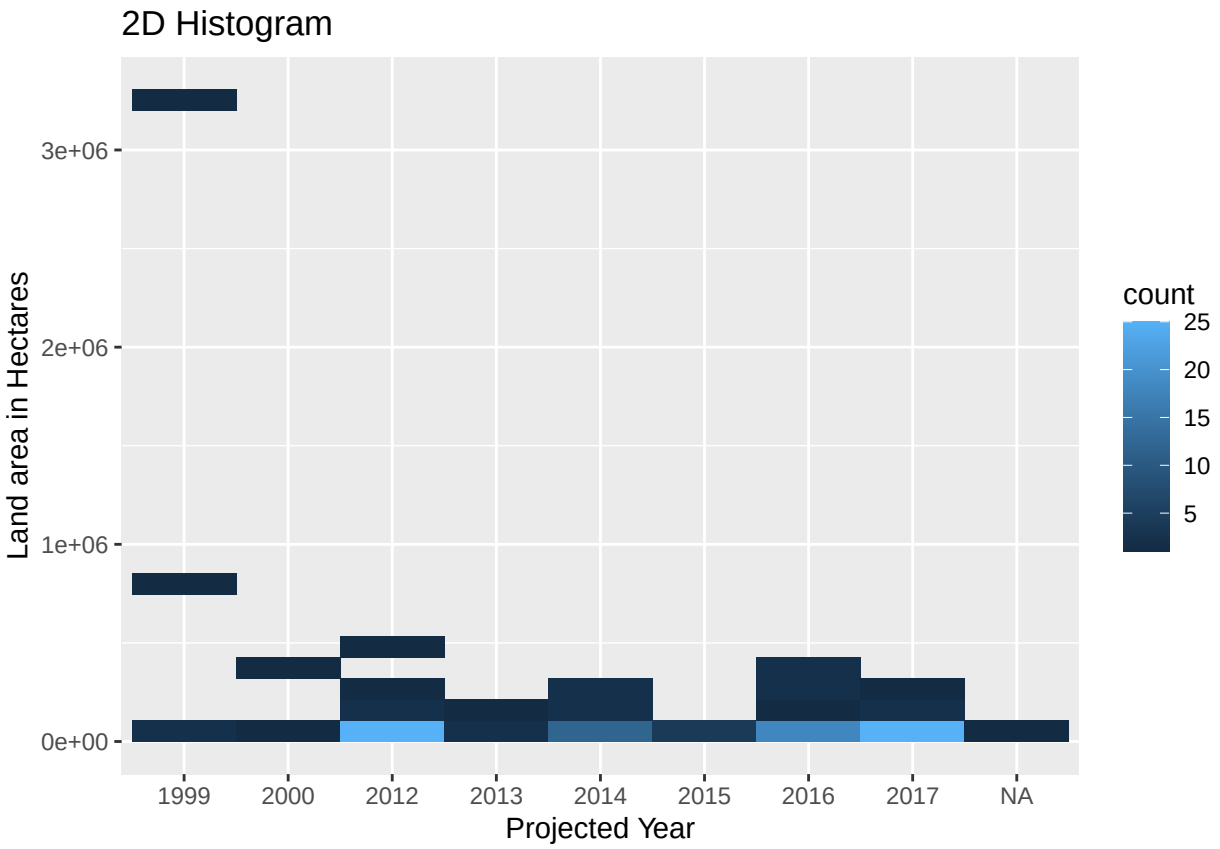
```
## Warning: NAs introduced by coercion
```

```
plot(combined_data$Hectares,
     combined_data$`Projected investment`,
     main = "Scatterplot of Hectares vs. Projected Investment",
     xlab = "Hectares",
     ylab = "Projected Investment",
     pch = 19,
     col = "blue")
```




```
#
# A 2d histogram represents the hectares of land sales year wise
#
```

```r
ggplot(combined_data, aes(x = Year, y = Hectares)) +
  geom_bin2d(bins = 30) +
  labs(
    title = "2D Histogram",
    x = "Projected Year",
    y = "Land area in Hectares"
  )
```



Writing the cleaned data set named “filter_sheet1”, “filter_sheet2” into an new excel file

```r
wb <- createWorkbook()

# We have filtered data sheets  "filter_sheet1", "filter_sheet2"

# Adding each filtered sheet into the new Excel file
addWorksheet(wb, sheetName = "Sheet1")
writeData(wb, sheet = "Sheet1", filter_sheet1)

addWorksheet(wb, sheetName = "Sheet2")
writeData(wb, sheet = "Sheet2", filter_sheet2)

# Save the Excel workbook to a file
saveWorkbook(wb, file = "filtered_data_Mounika.xlsx")
```