FloresDSC520Step2

Emilio Flores

2024-07-30

How to import and clean my data

- Almond Historical Production and Acreage Data:

```
# Import Excel files using the readxl library
library(readxl)

almond_df <- read_excel(
    paste0("C:/Users/emili/OneDrive - Bellevue University/Summer 2024/",
           "DSC 520 Statistics for Data Science/Final Project/",
           "AlmondData.xlsx"))

# Simplify column titles
colnames(almond_df) <- c("Year", "Bearing_Acr", "Yield/AC_(Lbs.)",
                         "Production(Mill_Lbs.)", "Price(c/lb.", "Value($1000")
```

- Fresno Area - Historical Precipitation Data:

```
# Import Fresno Area annual precipitation (top almond producer area)

fresno_precipation <- read_excel(
    paste0("C:/Users/emili/OneDrive - Bellevue University/Summer 2024/",
           "DSC 520 Statistics for Data Science/Final Project/",
           "FresnoPrecipitation.xlsx"))
```

- Bakersfield Area - Historical Precipitaion Data:

```
# Import Bakersfield Area annual precipitation (top almond producer area)

bakersfield_precipation <- read_excel(
    paste0("C:/Users/emili/OneDrive - Bellevue University/Summer 2024/",
           "DSC 520 Statistics for Data Science/Final Project/",
           "BakersfieldPrecipitation.xlsx"))
```

- Merge Relevant Columns Into New Data Set:

```
# Merged data sets by extracting specifc columns from different data sets into a
# new data frame

clean_df <- data.frame(
    Year = almond_df$Year,
    Production_MillLbs = almond_df$`Production(Mill_Lbs.)`,
    Acreage_1000s = almond_df$Bearing_Acr,
    Fresno_Prec_in = fresno_precipation$Annual,
    Bakersfield_Prec_in = bakersfield_precipation$Annual
)
```

What does the final data set look like?

```
head(clean_df)
```

```
##   Year Production_MillLbs Acreage_1000s Fresno_Prec_in Bakersfield_Prec_in
## 1 1980                322         326.8          10.28                6.01
## 2 1981                408         326.2          10.01                6.07
## 3 1982                347         339.0          16.08                8.03
## 4 1983                242         360.0          21.61               10.86
## 5 1984                590         381.0           6.77                3.42
## 6 1985                465         409.0           8.40                4.26
```

```
summary(clean_df)
```

```
##       Year      Production_MillLbs Acreage_1000s    Fresno_Prec_in
##  Min.   :1980   Min.   : 242       Min.   : 326.2   Length:43
##  1st Qu.:1990   1st Qu.: 515       1st Qu.: 414.5   Class :character
##  Median :2001   Median : 833       Median : 530.0   Mode  :character
##  Mean   :2001   Mean   :1177       Mean   : 638.3
##  3rd Qu.:2012   3rd Qu.:1880       3rd Qu.: 810.0
##  Max.   :2022   Max.   :3115       Max.   :1350.0
##  Bakersfield_Prec_in
##  Length:43
##  Class :character
##  Mode  :character
##
##
##
```

Questions for future steps.

- I can import data sets relatively easy but I lack on cleaning skills. Several of the Excel files that contained the data that I used in my final project had small footnotes and/or subtitles. R analyzed that additional text as rows of data and/or as columns. It would have been easier for me to clean the data in Excel than using R, specially because my data sets are not that large. I see how this would be an issue when working with data sets that have thousands of rows and columns.

What information is not self-evident?

- I originally thought that historical precipitation data would be sufficient to create a model that could predict almond production in California. After further analysis, I realized that precipitation alone is not the only factor affecting production. For example, total production at anytime will be affected by the number of planted acres in California. Running different regression models will give me a better idea of what variables may have a greater impact on production.

What are different ways you could look at this data?

- My original hypothesis was that precipitation can determine total almond production. I believe that this might not be the most accurate assumption as there are other variables that could determine total production, for example, bearing acreage. I could also try to test how almond yields per acre change according to precipitation levels. This could make it easier to understand how, regardless of planted acreage, almond tree productivity is affected by drought and/or other weather events.

How do you plan to slice and dice the data?

- I currently have an appropriate data set to test my hypothesis. If my original hypothesis and/or models don't generate significant results, I would refine my scope of work. For example, I am currently assuming that the precipitation data of the Fresno and Bakersfield areas would be enough to predict almond production because most of the almond trees in CA are planted near those cities. I could rather research the specific historical almond production in Fresno county and compare it against only the historical precipitation of that same area.

How could you summarize your data to answer key questions?

- I could summarize my data by calculating the average precipitation levels in the region. I could also calculate annual growth of almond production and/or the average yields per acre. Also, after running linear regressions, I could plot the relationship between rain and production levels.

What types of plots and tables will help you to illustrate the findings to your questions?

- Histograms will be beneficial to understand the distribution of production and/or yields per acre. This would allow me to understand how much yields have moved away from the mean due to intense drought or heavy rains.

- Scatter plots will also help me visualize the relationship between production and precipitation levels, if any.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

- I will use a logistical regression and us the predict function to measure the accuracy of my model. This might generate better insight on what variables have the greatest impact on my model(s).

Questions for future steps.

- I will explore the possibility to refine my scope of work to reduce the size of my data frame.
- I will research other plots and/or graphs that could visualize the trends in my data sets in a more efficient way.