Beans

Surabhi Metpally

2024-10-29

Prescriptive Analysis of Bean Data

In this section, I will conduct a prescriptive analysis of the bean dataset provided. The goal is to preprocess the data by cleaning it, removing unnecessary columns, and filtering for specific classes of beans. I will also create new variables that will help in categorizing the data based on convex area.

Following the data transformation, I will perform statistical summaries to gain insights into the characteristics of different bean classes. This will involve calculating measures such as the median area, standard deviation, and percentiles for key variables. Additionally, I will construct a contingency table to explore the relationship between different classifications of beans and their convex area categories. This analysis aims to provide a comprehensive overview of the data, helping to identify patterns and make informed decisions regarding the bean varieties examined.

```
# Load the necessary libraries
library(tidyverse)
library(readxl)

# Read the data
beans_data <- read_excel("/Users/surabhimetpally/Downloads/smetpal_beans.xlsx")

# Check the first few rows of the data
head(beans_data)
```

```
## # A tibble: 6 x 17
##     Area Perimeter MajorAxisLength MinorAxisLength AspectRatio Eccentricity
##    <dbl>     <dbl>           <dbl>           <dbl>       <dbl>        <dbl>
## 1  31719      658.            241.            168.        1.43        0.715
## 2  40647      738.            248.            209.        1.18        0.535
## 3 204023     1700.            622.            422.        1.47        0.734
## 4  37790      711.            259.            187.        1.39        0.692
## 5  43924      834.            345.            163.        2.12        0.882
## 6  24209      579.            220.            141.        1.56        0.769
## # i 11 more variables: ConvexArea <dbl>, EquivDiameter <dbl>, Extent <dbl>,
## #   Solidity <dbl>, Roundness <dbl>, Compactness <dbl>, ShapeFactor1 <dbl>,
## #   ShapeFactor2 <dbl>, ShapeFactor3 <dbl>, ShapeFactor4 <dbl>, Class <chr>
```

```
# Data Cleaning and Transformation: Select, Rename, Filter, Create New Variables, and Reorder

# Remove unwanted columns and rename
cleaned_data <- beans_data %>%
  select(-Roundness, -Solidity) %>%
```

```
  rename(SF1 = ShapeFactor1, SF2 = ShapeFactor2, SF3 = ShapeFactor3, SF4 = ShapeFactor4)

# Filter by class
filtered_data <- cleaned_data %>%
  filter(Class %in% c("DERMASON", "BARBUNYA", "HOROZ", "CALI", "SIRA"))

# Create new variables
transformed_data <- filtered_data %>%
  mutate(Avg_SF = rowMeans(select(., starts_with("SF")))) %>%
  mutate(ConvexArea_Category = case_when(
    ConvexArea > 51625.6 ~ "largest",
    ConvexArea > 38439.4 & ConvexArea <= 51625.6 ~ "middle",
    TRUE ~ "lowest"
  ))

# Reorder data
final_data <- transformed_data %>%
  arrange(desc(Eccentricity))

# Display the final data
final_data
```

```
## # A tibble: 7,717 x 17
##     Area Perimeter MajorAxisLength MinorAxisLength AspectRatio Eccentricity
##    <dbl>     <dbl>           <dbl>           <dbl>       <dbl>        <dbl>
##  1 51511      936.            397.            166.        2.39        0.908
##  2 58410      983.            420.            178.        2.36        0.906
##  3 54508      968.            404.            172.        2.35        0.905
##  4 50107      913.            387.            165.        2.34        0.904
##  5 56849      989.            412.            176.        2.33        0.904
##  6 59651     1029.            422.            181.        2.33        0.903
##  7 64919     1032.            439.            188.        2.33        0.903
##  8 48701      900.            381.            164.        2.33        0.903
##  9 58869     1011.            418.            180.        2.32        0.902
## 10 49815      910.            384.            166.        2.32        0.902
## # i 7,707 more rows
## # i 11 more variables: ConvexArea <dbl>, EquivDiameter <dbl>, Extent <dbl>,
## #   Compactness <dbl>, SF1 <dbl>, SF2 <dbl>, SF3 <dbl>, SF4 <dbl>, Class <chr>,
## #   Avg_SF <dbl>, ConvexArea_Category <chr>
```

```
# Contingency Table of Bean Class by Convex Area Category for SF1 < 0.0081

# Filter data for SF1 < 0.0081
filtered_data <- final_data %>%
  filter(SF1 < 0.0081)

# Create a two-way contingency table
contingency_table <- table(filtered_data$Class, filtered_data$ConvexArea_Category)

# Print the table
contingency_table
```

```
##
```

```
##            largest lowest middle
##   BARBUNYA     897      0     37
##   CALI        1123      0      1
##   DERMASON       0   1492    320
##   HOROZ        917     11    420
##   SIRA         150    121   1520
```

```r
# Summary Statistics for Area and Perimeter by Class

library(dplyr)

bean_stats <- final_data %>%
  group_by(Class) %>%
  summarize(
    Median_Area = median(Area),
    SD_Area = sd(Area),
    Percentile_95_Area = quantile(Area, 0.95),
    Min_Area = min(Area),
    Median_Perimeter = median(Perimeter),
    SD_Perimeter = sd(Perimeter),
    Percentile_95_Perimeter = quantile(Perimeter, 0.95),
    Min_Perimeter = min(Perimeter)
  )

print(bean_stats)
```

```
## # A tibble: 5 x 9
##   Class    Median_Area SD_Area Percentile_95_Area Min_Area Median_Perimeter
##   <chr>          <dbl>   <dbl>              <dbl>    <dbl>            <dbl>
## 1 BARBUNYA      69234   10458.             87032.    41487            1042.
## 2 CALI          74870.   9544.             92828.    45666            1056.
## 3 DERMASON      31958    4673.             40011.    20464             665.
## 4 HOROZ         53715    7351.             65476     33006             922.
## 5 SIRA          44568    4546.             52162.    31519             794.
## # i 3 more variables: SD_Perimeter <dbl>, Percentile_95_Perimeter <dbl>,
## #   Min_Perimeter <dbl>
```