Fertilizer Dataset Exploration

David Goh

2025-09-10

Load Data

```
train <- read.csv("playground-series-s5e6/train.csv")
test <- read.csv("playground-series-s5e6/test.csv")
str(train)
```

```
## 'data.frame':    750000 obs. of  10 variables:
##  $ id             : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Temparature    : int  37 27 29 35 35 30 27 36 36 28 ...
##  $ Humidity       : int  70 69 63 62 58 59 62 62 51 50 ...
##  $ Moisture       : int  36 65 32 54 43 29 53 44 32 35 ...
##  $ Soil.Type      : chr  "Clayey" "Sandy" "Sandy" "Sandy" ...
##  $ Crop.Type      : chr  "Sugarcane" "Millets" "Millets" "Barley" ...
##  $ Nitrogen       : int  36 30 24 39 37 10 26 30 19 25 ...
##  $ Potassium      : int  4 6 12 12 2 0 15 12 17 12 ...
##  $ Phosphorous    : int  5 18 16 4 16 9 22 35 29 16 ...
##  $ Fertilizer.Name: chr  "28-28" "28-28" "17-17-17" "10-26-26" ...
```

```
colnames(train)
```

```
##  [1] "id"              "Temparature"     "Humidity"        "Moisture"
##  [5] "Soil.Type"       "Crop.Type"       "Nitrogen"        "Potassium"
##  [9] "Phosphorous"     "Fertilizer.Name"
```

```
dim(train)
```

```
## [1] 750000     10
```

```
## dim: 750000 x 10
```

```
## any NA anywhere: FALSE
```

```
##              id     Temperature        Humidity        Moisture       Soil.Type
##               0               0               0               0               0
##       Crop.Type        Nitrogen       Potassium     Phosphorous Fertilizer.Name
##               0               0               0               0               0
```

```
## 
## unique counts (Soil, Crop, Fertilizer):
```

```
##       Soil.Type       Crop.Type Fertilizer.Name
##               5              11               7
```

```
## 
## class counts for target:
```

```
## 
## 14-35-14 10-26-26 17-17-17    28-28    20-20      DAP     Urea
##   114436   113887   112453   111158   110889    94860    92317
##
## any duplicated id?: FALSE
```

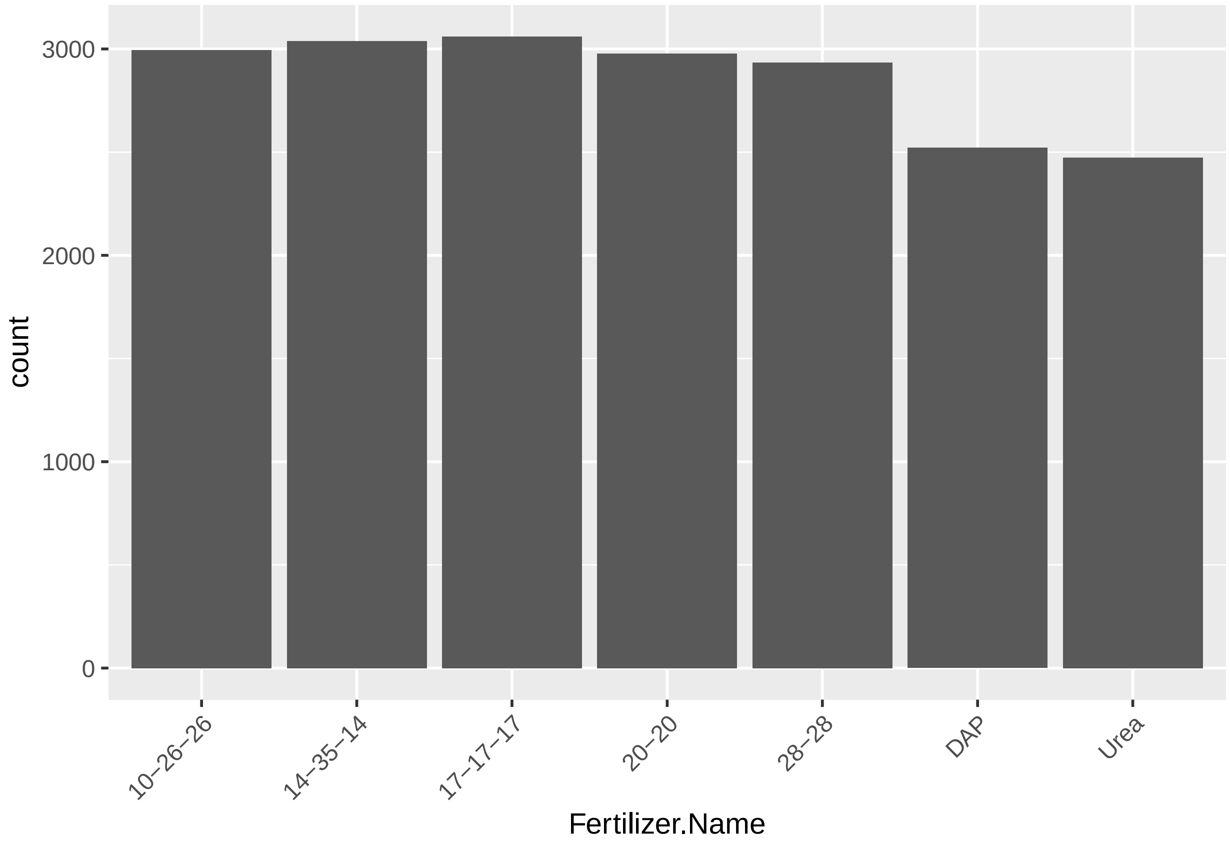
```r
nums <- select_if(train, is.numeric) %>% select(-id)
print(summary(nums))
```

```
##   Temperature      Humidity       Moisture       Nitrogen
##  Min.   :25.0   Min.   :50.00   Min.   :25.00   Min.   : 4.00
##  1st Qu.:28.0   1st Qu.:55.00   1st Qu.:35.00   1st Qu.:13.00
##  Median :32.0   Median :61.00   Median :45.00   Median :23.00
##  Mean   :31.5   Mean   :61.04   Mean   :45.18   Mean   :23.09
##  3rd Qu.:35.0   3rd Qu.:67.00   3rd Qu.:55.00   3rd Qu.:33.00
##  Max.   :38.0   Max.   :72.00   Max.   :65.00   Max.   :42.00
##    Potassium        Phosphorous
##  Min.   : 0.000   Min.   : 0.00
##  1st Qu.: 4.000   1st Qu.:10.00
##  Median : 9.000   Median :21.00
##  Mean   : 9.478   Mean   :21.07
##  3rd Qu.:14.000   3rd Qu.:32.00
##  Max.   :19.000   Max.   :42.00
```

```r
print(sapply(nums, function(x) length(unique(x))))
```

```
## Temperature    Humidity    Moisture    Nitrogen   Potassium Phosphorous
##          14          23          41          39          20          43
```

```r
set.seed(123); samp <- slice_sample(train, n = 20000)
ggplot(samp, aes(x = Fertilizer.Name)) + geom_bar() + theme(axis.text.x = element_text(angle=45, hjust=
```


3000
2000
1000
0
count
10–26–26
14–35–14
17–17–17
20–20
28–28
DAP
Urea
Fertilizer.Name