

Sorghum Yield Analysis Report

Virada Nan

2026-01-21

Table of Contents

- Download library
- Download Dataset
- R charts:
 1. Top 5 High-Yield Genotypes
 2. Bottom 5 Low-Yield Genotypes
 3. Yield Stability
 4. Yearly Yield Trends by Location
 5. Genotype Comparison

Download library

```
library(dplyr)
library(data.table)
library(agridat)
library(ggplot2)
library(tinytex)
```

Download Dataset

Select dataset

```
data("adugna.sorghum")
sorghum_yield <- adugna.sorghum
```

Manage data within dataset

```
sorghum_yield <- sorghum_yield %>%
  rename(genotype = gen,
         environment = env,
         location = loc)
```

Check data after revise names

```
glimpse(sorghum_yield, width = 60)
```

```
## Rows: 289
## Columns: 6
## $ genotype    <fct> G16, G17, G18, G19, G20, G21, G22, G23~
## $ trial       <fct> T2, T2, T2, T2, T2, T2, T2, T2, T2~
## $ environment <fct> E01, E01, E01, E01, E01, E01, E01, E01~
## $ yield       <int> 590, 554, 586, 738, 489, 684, 555, 102~
## $ year        <int> 2001, 2001, 2001, 2001, 2001, 2001, 20~
## $ location    <fct> Mieso, Mieso, Mieso, Mieso, Mieso, Mie~
```

```
head(sorghum_yield)
```

```
##   genotype trial environment yield year location
## 1     G16     T2          E01   590 2001    Mieso
## 2     G17     T2          E01   554 2001    Mieso
## 3     G18     T2          E01   586 2001    Mieso
## 4     G19     T2          E01   738 2001    Mieso
## 5     G20     T2          E01   489 2001    Mieso
## 6     G21     T2          E01   684 2001    Mieso
```

R Chart

— 1. Top 5 High-Yield Genotypes —

Data Transformation

```
top_5_gen_data <- sorghum_yield %>%
  select(genotype, year, yield) %>%
  filter(year %in% c(2003, 2004, 2005)) %>%
  group_by(genotype) %>%
  summarise(avg_yield = round(mean(yield), 2), .groups = "drop") %>%
  arrange(desc(avg_yield)) %>%
  head(5)

print(top_5_gen_data)
```

```
## # A tibble: 5 x 2
##   genotype avg_yield
##   <fct>      <dbl>
## 1 G01        5080.
## 2 G03        5044.
## 3 G11        4950.
## 4 G04        4842
## 5 G06        4770.
```

Note

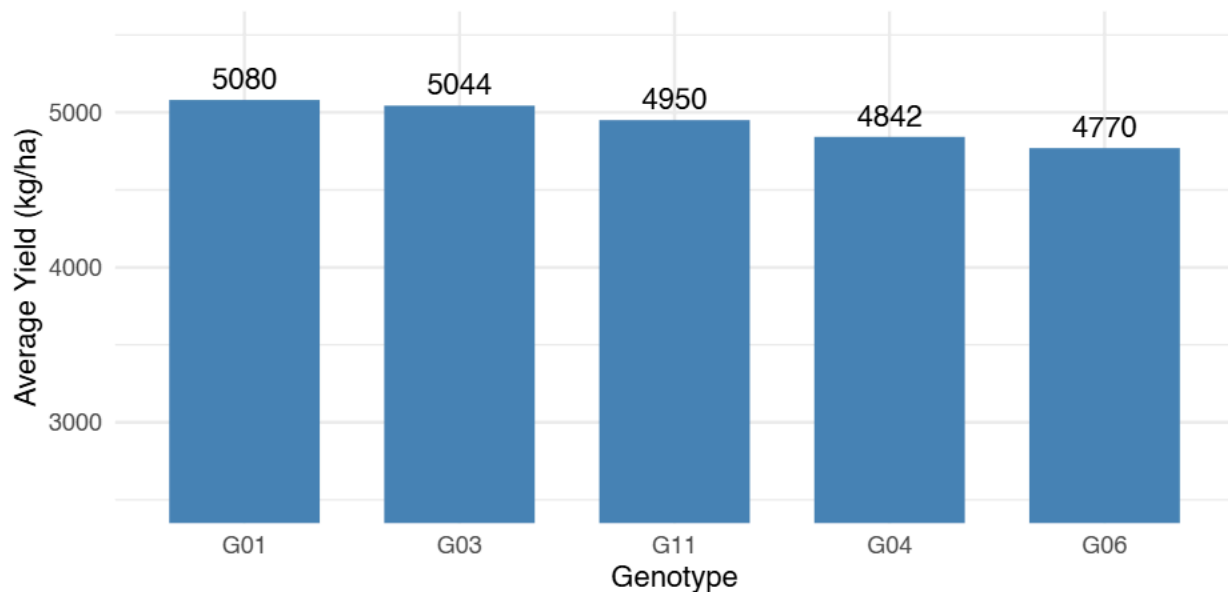
I filtered the data to include only the years 2003–2005. This ensures a fair comparison because all varieties were consistently cultivated during this specific period.

Data Visualization

```
ggplot(top_5_gen_data,
       aes(x = reorder(genotype, -avg_yield),
           y = avg_yield)) +
  geom_col(fill = "steelblue",
           width = 0.7) +
  geom_text(aes(label = round(avg_yield, 0)),
            vjust = -0.5,
            size = 4) +
  coord_cartesian(ylim = c(2500, 5500)) +
  theme_minimal() +
  labs(title = "Top 5 Genotypes by Average Yield",
       subtitle = "Annual data from 2003 - 2005",
       x = "Genotype",
       y = "Average Yield (kg/ha)",
       caption = "Data source: adugna.sorghum {agridat} dataset in R
                 \nY-axis range: 2,500 - 5,500 for better comparison
                 \nThis data includes only the last 3 years (2003 - 2005)
                 because all varieties were consistently cultivated during this period.")
```

Top 5 Genotypes by Average Yield

Annual data from 2003 – 2005



Data source: `adugna.sorghum {agridat}` dataset in R

Y-axis range: 2,500 – 5,500 for better comparison

This data includes only the last 3 years (2003 – 2005) because all varieties were consistently cultivated during this period.

Summary

This chart identifies the top five varieties with the highest average yield. These findings provide a data-driven basis for recommending the most productive varieties for large-scale farming.

— 2. Bottom 5 Low-Yield Genotypes —

Data Transformation

```
bottom_5_gen_data <- sorghum_yield %>%
  select(genotype, year, yield) %>%
  filter(year %in% c(2003, 2004, 2005)) %>%
  group_by(genotype) %>%
  summarise(avg_yield = round(mean(yield), 2), .groups = "drop") %>%
  arrange(avg_yield) %>%
  head(5)

print(bottom_5_gen_data)
```

```
## # A tibble: 5 x 2
##   genotype avg_yield
##   <fct>      <dbl>
## 1 G26        3396.
## 2 G27        3572.
## 3 G17        3650.
## 4 G25        3657.
## 5 G16        3660.
```

Note

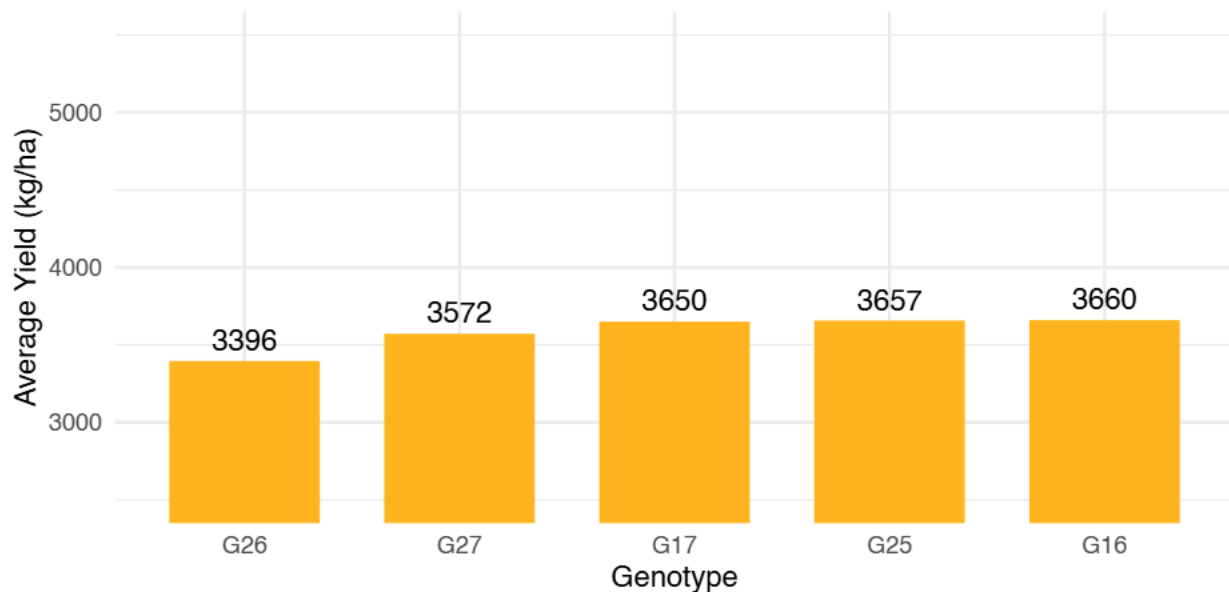
I select the data to include only the years 2003-2005 because all varieties were consistently cultivated during this period, ensuring a fair comparison.

Data Visualization

```
ggplot(bottom_5_gen_data,
  aes(x = reorder(genotype, avg_yield),
    y = avg_yield)) +
  geom_col(fill = "#FFB823",
    width = 0.7) +
  geom_text(aes(label = round(avg_yield, 0)),
    vjust = -0.5,
    size = 4) +
  coord_cartesian(ylim = c(2500, 5500)) +
  theme_minimal() +
  labs(title = "Bottom 5 Genotypes by Average Yield",
    subtitle = "Annual data from 2003 - 2005",
    x = "Genotype",
    y = "Average Yield (kg/ha)",
    caption = "Data source: adugna.sorghum {agridat} dataset in R
      \nY-axis range: 2,500 - 5,500 for better comparison
      \nThis data includes only the last 3 years (2003 - 2005)
      because all varieties were consistently cultivated during this period.")
```

Bottom 5 Genotypes by Average Yield

Annual data from 2003 – 2005



Data source: `adugna.sorghum {agridat}` dataset in R

Y-axis range: 2,500 – 5,500 for better comparison

This data includes only the last 3 years (2003 – 2005) because all varieties were consistently cultivated during this period.

Summary

This chart identifies the five varieties with the lowest average yield. These results help decision-makers consider excluding them from future planting plans to reduce farming costs.

— 3. Yield Stability —

Data Transformation

- Preparation
- Summary Table

Preparation

```
boxplot_data <- sorghum_yield %>%  
  select(genotype, year, yield) %>%  
  filter(year %in% c(2003, 2004, 2005)) %>%  
  filter(genotype %in% c("G01", "G03", "G11", "G04", "G06"))  
  
head(boxplot_data, 3)
```

```
##   genotype year yield  
## 1      G01 2003  4800  
## 2      G03 2003  4700  
## 3      G04 2003  4650
```

Note

I select the “Top 5 Genotypes” based on their average yield to further observe and compare their yield stability.

Summary Table

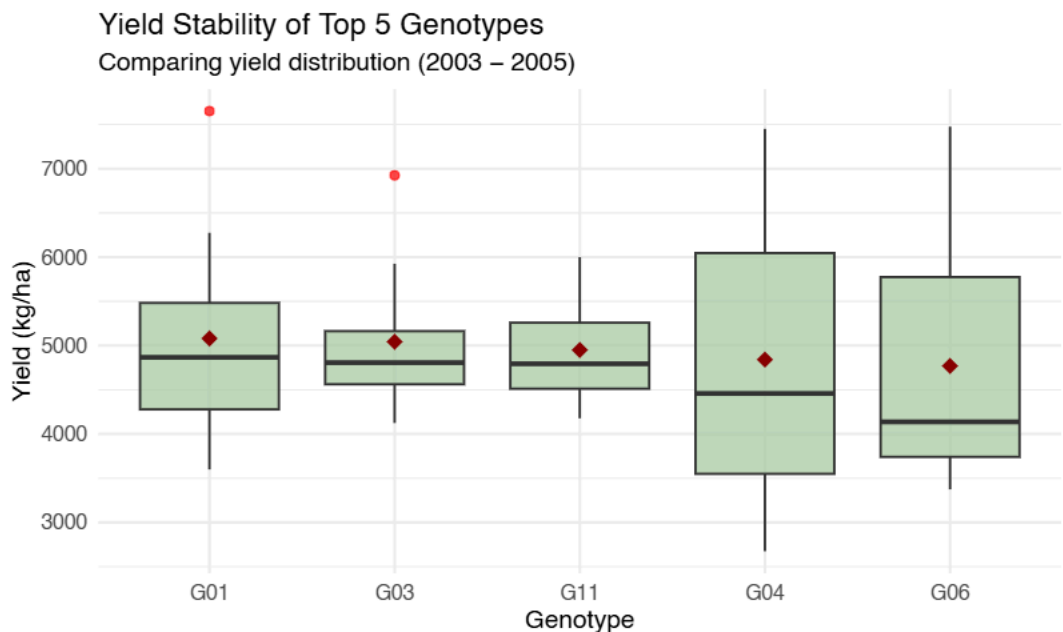
```
top_5_stability_check <- boxplot_data %>%  
  group_by(genotype) %>%  
  summarise(  
    avg_yield = mean(yield),  
    median_yield = median(yield),  
    avg_med_diff = avg_yield - median_yield,  
    min_yield = min(yield),  
    max_yield = max(yield),  
    sd_yield = sd(yield),  
    .groups = "drop"  
  ) %>%  
  arrange(desc(avg_yield))  
  
knitr::kable(top_5_stability_check,  
  caption = "Stability Statistics of Top 5 Genotypes",  
  digits = 2)
```

Table 1: Stability Statistics of Top 5 Genotypes

genotype	avg_yield	median_yield	avg_med_diff	min_yield	max_yield	sd_yield
G01	5080.12	4867.0	213.12	3600	7650	1343.93
G03	5043.50	4807.0	236.50	4125	6925	942.05
G11	4950.12	4794.0	156.12	4178	6000	661.25
G04	4842.00	4458.5	383.50	2675	7450	1764.98
G06	4769.50	4137.5	632.00	3375	7475	1453.14

Data Visualization

```
ggplot(boxplot_data,
       mapping = aes(x = reorder(genotype, -yield),
                      y = yield)) +
  geom_boxplot(fill = "#A5C89E",
               outlier.color = "red",
               alpha = 0.7) +
  stat_summary(fun = mean, geom = "point", shape = 18, size = 3, color = "darkred") +
  theme_minimal() +
  labs(title = "Yield Stability of Top 5 Genotypes",
       subtitle = "Comparing yield distribution (2003 - 2005)",
       x = "Genotype",
       y = "Yield (kg/ha)",
       caption = "Data source: adugna.sorghum {agridat} dataset in R
                 \n The box represents the yield distribution.
                 \n The diamond shape indicates the mean")
```



Data source: adugna.sorghum {agridat} dataset in R

The box represents the yield distribution.
The diamond shape indicates the mean

Summary

- **Most Reliable:** G11 is the most stable genotype. It has the shortest overall range (whiskers) and the lowest SD (661), meaning the yield is highly consistent.
- **High Potential with Risk:** G01 and G03 have the highest average yields. They have high outliers (nearly 8,000 kg/ha), showing great potential, but they have more variation than G11.
- **Low Consistency:** G04 and G06 show high variance. Their `avg_med_diff` is very high (up to 632), and their boxes are the largest, which means their yield is unpredictable.
- **Business Recommendation:** We should recommend G11 for farmers who want steady income, and G01 for those who want to reach the highest possible yield.

— 4. Yearly Yield Trends by Location —

Data Transformation

```
location_trend <- sorghum_yield %>%
  select(location, year, yield) %>%
  group_by(location, year) %>%
  summarise(avg_yield = mean(yield), .groups = "drop")

head(location_trend)
```

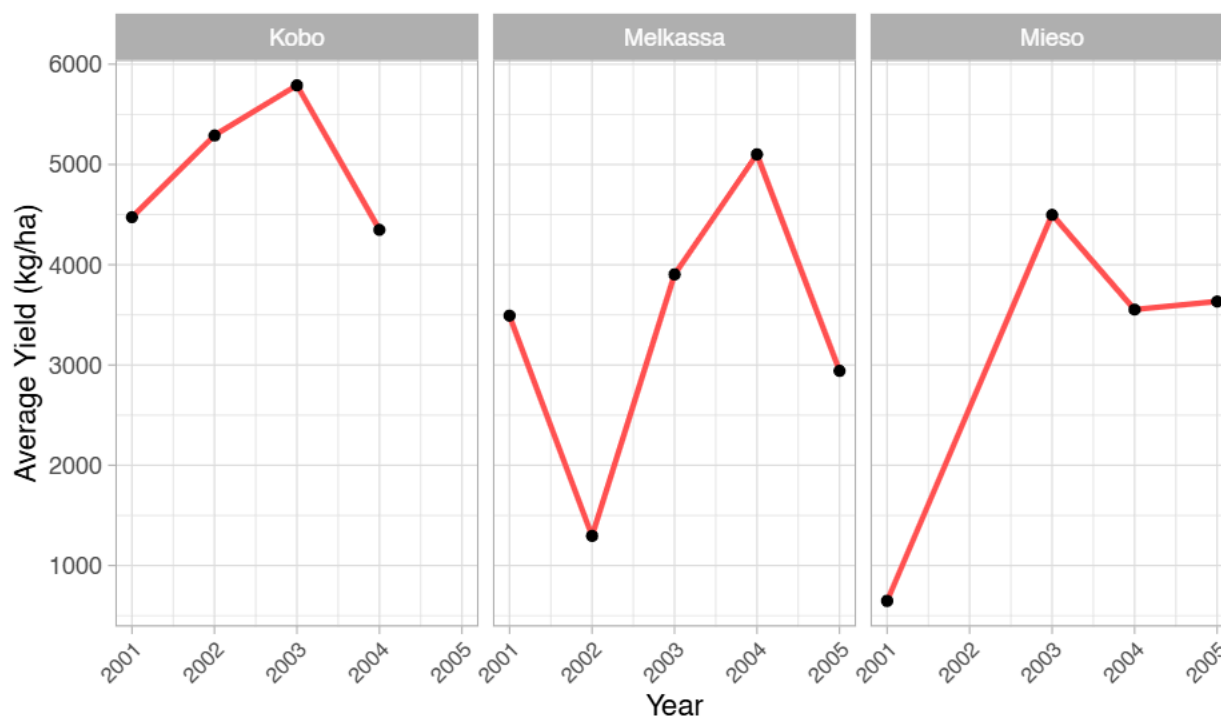
```
## # A tibble: 6 x 3
##   location  year avg_yield
##   <fct>    <int>   <dbl>
## 1 Kobo      2001    4474
## 2 Kobo      2002    5290.
## 3 Kobo      2003    5789.
## 4 Kobo      2004    4349.
## 5 Melkassa 2001    3492.
## 6 Melkassa 2002    1296.
```

Data Visualization

```
ggplot(location_trend,
  mapping = aes(x = year,
                y = avg_yield)) +
  geom_line(color = "#FF5555",
    linewidth = 1) +
  geom_point() +
  facet_wrap(~location, ncol = 3) +
  theme_light() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1, size = 8)) +
  labs(title = "Annual Yield Trends by Location",
    subtitle = "Performance comparison yield by location (2001 - 2005)",
    x = "Year",
    y = "Average Yield (kg/ha)",
    caption = "Data source: adugna.sorghum {agridat} dataset in R")
```

Annual Yield Trends by Location

Performance comparison yield by location (2001 – 2005)



Data source: adugna.sorghum {agridat} dataset in R

Summary

- **Best Location:** Kobo is the top performer. It reached the highest yield of nearly 6,000 kg/ha in 2003.
- **The 2003 Peak:** All three locations showed a significant increase (jump) in yield in 2003. This happened when new genotypes (G01–G15) were introduced for testing.
- **High Volatility:** Melkassa is the least stable location. It had a deep drop in 2002 before recovering later.
- **Note:** This analysis focuses on location potential and genotypes. External factors like weather or soil data were not available.

— 5. Genotype Comparison —

Data Transformation

```
comp_g01_g11 <- sorghum_yield %>%  
  select(genotype, year, location, yield) %>%  
  filter(genotype %in% c("G01", "G11")) %>%  
  group_by(location, genotype, year) %>%  
  summarise(avg_yield = mean(yield), .groups = "drop")  
  
head(comp_g01_g11)
```

```
## # A tibble: 6 x 4  
##   location genotype  year avg_yield  
##   <fct>      <fct>   <int>    <dbl>  
## 1 Kobo      G01      2003     7650  
## 2 Kobo      G01      2004     5218  
## 3 Kobo      G11      2003     6000  
## 4 Kobo      G11      2004     5810  
## 5 Melkassa G01      2003     4475  
## 6 Melkassa G01      2004     6275
```

Note

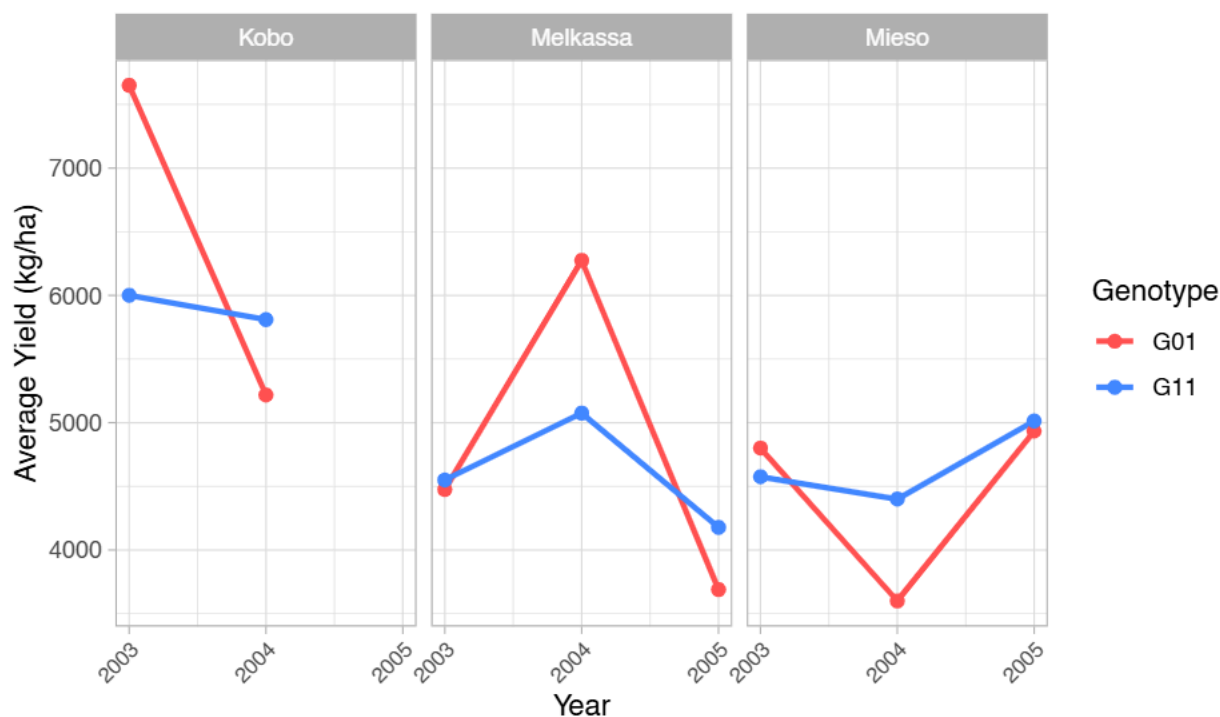
I focused on G01 and G11 only to compare their performance. G01 represents the highest-yield potential, while G11 represents the best yield stability.

Data Visualization

```
ggplot(comp_g01_g11,  
  mapping = aes(x = year,  
                y = avg_yield,  
                color = genotype,  
                group = genotype)) +  
  geom_line(linewidth = 1) +  
  geom_point(size = 2) +  
  facet_wrap(~location) +  
  scale_color_manual(values = c("G01" = "#FF5555", "G11" = "#4488FF")) +  
  scale_x_continuous(breaks = 2003:2005) +  
  theme_light() +  
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1, size = 8)) +  
  labs(title = "Genotype Comparison: G01 vs G11",  
       subtitle = "Comparing High-Yield (G01) vs High-Stability (G11) across locations",  
       x = "Year",  
       y = "Average Yield (kg/ha)",  
       caption = "Data source: adugna.sorghum {agridat} dataset in R",  
       color = "Genotype")
```

Genotype Comparison: G01 vs G11

Comparing High-Yield (G01) vs High-Stability (G11) across locations



Data source: adugna.sorghum {agridat} dataset in R

Summary

- **Yield vs. Stability:** G01 shows high volatility (sharp upward and downward movements) in all three locations. While G11 is much more stable than G01.
- **Location Recommendation (Mieso):** For cultivation in Mieso, G11 is highly recommended over G01. The Mieso graph shows that G11 performs better overall and remains more consistent throughout the years.
- **G01 (High Risk, High Reward):** G01 is suitable for farmers who want the highest possible yield, but they must accept high risk from its unpredictability.
- **G11 (Safe & Steady):** G11 is suitable for farmers who prefer a consistent income. It provides a steady yield even when conditions change.
- **Final Choice:** If the priority is maximum profit, choose G01. If the priority is low risk, choose G11.