SOUTH DAKOTA STATE UNIVERSITY

STATISTICAL METHODS II (STAT-541)

Wheat Genotypic Performance Under Variable Nitrogen Fertilization: A Comparative Study Across Two Locations

Group: AAHS

Contributors: Amool Singh Vadithya, Ashish Chaudhary, Horender Sharma, Shivam Singh

Instructor: Dr. Gemechis Djira
GTA: Mr. Iftekhar Chowdhury
Semester: Fall 2025

The following document was developed by Amool Singh Vadithya, Ashish Chaudhary, Horender Sharma, and Shivam Singh, students at South Dakota State University, as a requisite for the course STAT 541 (Statistical methods II).

Contributor	Section(s)
Amool Singh Vadithya	Final Report Compilation, Statistical Analyses, Results, Conclusion
Ashish Chaudhary	Final Report Compilation, Introduction, Materials & Methods
Horender Sharma	Final Report Compilation, Statistical Analyses, Results, Conclusion
Shivam Singh	Final Report Compilation, Introduction, Materials & Methods
All team members	Report Review

Content

1. Introduction

1.1 General Introduction or Problem Statement

Winter wheat (*Triticum aestivum* L.) is a global staple and a cornerstone of agricultural economics, particularly in the U.S. Great Plains and regions like South Dakota (SD). The crop is vital for **global food security**, providing essential energy, protein, and micronutrients to billions of people. The yield and quality of winter wheat are overwhelmingly governed by the application of Nitrogen (N), the single most significant manageable input (Zhang et al., 2021). However, N fertilization presents a critical economic and environmental dilemma:

a. Environmental Cost: Excessive N use is a primary driver of nitrate leaching, water contamination, and increased greenhouse gas emissions, directly contradicting goals for agricultural sustainability (Cassman et al., 2002).

b. Economic Cost: Insufficient N application compromises grain yield and protein content, leading to substantial financial losses for farmers.

The challenge lies in determining the optimal N rate, which is not a constant value but is highly dependent on the complex Genotype x Environment x Nitrogen (G x E x N) interaction. Winter wheat breeding has created genotypes with distinct physiological N-responses, and these responses are further modulated by specific environmental conditions, such as the differences found between SD's Aurora and Dakota Lakes research farms. Therefore, understanding this three-way interaction is essential for developing precision agriculture protocols that maximize Nitrogen Use Efficiency (NUE) while maintaining high yield potential (Parent et al., 2017). This project addresses this imperative by providing a robust, data-driven analysis of all the interactions between the components using 2025 field data, **with the goal of generating genotype- and location-specific N management** recommendations for South Dakota winter wheat production.

2. The Data

2.1 Dataset and Source Description

The dataset was generated from a 2025 field experiment conducted by the Wheat Breeding Lab at South Dakota State University (SDSU). The data were collected from two distinct South Dakota research farms-the Aurora Research Farm and the Dakota Lakes Research Farm-which represent two different environmental conditions within the state. The experiment employed a three-factor factorial treatment structure (Genotype x Nitrogen Rate x Location) arranged in a Randomized Complete Block Design (RCBD) with a split-plot design structure. This design was chosen for its

practicality in applying N treatments (main plots) and its efficiency in analyzing the critical interaction effects. The total experiment size is 240 observations (4 N-Doses x 10 Genotypes x 3 Replications x 2 Locations). The data are unique, as they have neither been explored nor published prior to this project.

2.2 Definition and Classification of Variables

The experiment evaluates the genotypic variability in response to N fertilization across different locations. The variables are classified as follows:

a) **Nitrogen Dose (Independent Variable)**
   - Levels: 40 lbs (Control), 80 lbs, 100 lbs, 130 lbs.
   - Type: Quantitative, discrete

b) **Genotypes (Independent Variable)**
   - Levels: 10 genotypes (WW1 to WW10).
   - Type: Categorical, nominal.

c) **Locations (Independent Variable)**
   - Levels: 2 distinct locations (Aurora and the Dakota Lake).
   - Type: Categorical, nominal.

d) **Traits (Dependent Variables):**
   - Yield, Grain Protein Content, Test Weight, Grain Nitrogen and Heading Date (DHD)
   - Type: Quantitative, continuous.

2.3 First 10 Observations (Illustrative SPSS Data View)

The table below provides a snapshot of the first 10 observations, demonstrating the format of the raw data.

Table 1: Overview of the data being used in current project.

	A	B	C	D	E	F	G	H	I
1	**Location**	**Replication**	**N_Dose**	**Genotypes**	**DHD**	**Grain_Yield**	**Protein_%**	**Test_Weight**	**Grain_N**
2	Aurora	R1	40lbs	WW1	147	2055.288576	14.3	59.8	51.13969035
3	Aurora	R1	40lbs	WW2	146	1762.151794	14.1	59.5	43.23263211
4	Aurora	R1	40lbs	WW3	147	1925.413754	14.4	58.6	48.40914121
5	Aurora	R1	40lbs	WW4	149	1365.301695	15.9	56	37.60054848
6	Aurora	R1	40lbs	WW5	147	3012.323066	10.9	59.3	57.46006248
7	Aurora	R1	40lbs	WW6	146	2515.088847	11.5	61.7	50.55805035
8	Aurora	R1	40lbs	WW7	149	2570.71631	11.1	62.3	49.76456239
9	Aurora	R1	40lbs	WW8	146	2796.713349	12.1	59.6	59.01689861
10	Aurora	R1	40lbs	WW9	145	2298.19047	13	59.3	52.10416366
11	Aurora	R1	40lbs	WW10	146	1989.721111	12.7	61.1	44.01916473

2.4 Objectives and Hypothesis

The overall objective of this project is to **investigate the effects of varying nitrogen application rates on wheat production across different genotypes at locations (Aurora and Dakota Lake)**, with the practical goal of **identifying and recommending the specific genotypes performing best under different nitrogen fertility levels and environmental conditions**.

The study tests the null hypothesis that nitrogen, genotype, location or their interactions have no significant effect on yield, against the alternative hypothesis that these factors or interactions have a statistically significant impact.

3. Descriptive analyses

3.1 Descriptive and Preliminary Analyses: Location-Wise Comparison

Descriptive statistics and graphical analyses were executed to characterize the data and provide preliminary insights into the main and interaction effects before formal inferential testing. The analysis focuses on comparative performance between the **Aurora Research Farm** and the **Dakota Lakes Research Farm**.

Table 2: Descriptive statistics for grain yield (kg/ha) data of 10 wheat genotypes.

Location	Genotypes	Mean	SD	CV %	Minimum	Maximum	Median
Aurora	WW1	2211.56	664.98	30.07	1041.68	3293.12	2273.08
Aurora	WW2	2237.07	688.59	30.78	1117.67	3190.21	2257.29
Aurora	WW3	2256.90	757.02	33.54	925.26	3121.24	2514.45
Aurora	WW4	2015.32	549.02	27.24	1169.54	2892.84	2194.79
Aurora	WW5	2624.75	555.73	21.17	1844.98	3570.02	2638.75
Aurora	WW6	2629.62	317.69	12.08	2104.33	3084.79	2623.09
Aurora	WW7	2712.79	438.67	16.17	1690.80	3438.76	2717.14
Aurora	WW8	2423.49	409.82	16.91	1762.15	2984.74	2439.25
Aurora	WW9	2301.43	417.13	18.12	1690.26	2989.37	2201.43
Aurora	WW10	2540.76	295.51	11.63	1989.72	3089.59	2568.22
Dakota Lakes	WW1	2616.56	372.30	14.23	2221.10	3647.41	2528.93
Dakota Lakes	WW2	2436.56	337.92	13.87	1781.96	3082.27	2487.31
Dakota Lakes	WW3	2621.93	291.49	11.12	2204.39	3082.55	2664.41
Dakota Lakes	WW4	2471.39	359.40	14.54	1740.69	3023.78	2472.92
Dakota Lakes	WW5	2456.19	300.07	12.22	2019.22	2971.61	2403.24
Dakota Lakes	WW6	2696.54	367.51	13.63	1799.39	3072.34	2808.30

Dakota Lakes	WW7	2604.55	300.02	11.52	2077.44	3005.00	2676.156088
Dakota Lakes	WW8	2352.55	310.48	13.20	1739.31	2638.82	2490.825116
Dakota Lakes	WW9	2353.67	414.32	17.60	1654.07	3001.83	2445.445068
Dakota Lakes	WW10	2551.32	271.78	10.65	2164.07	2919.61	2552.850195

Interpretation: Genotype **WW7** at Aurora showed the **highest mean yield** with a **lower CV value**, followed by **WW6** and **WW5**. At the Dakota Lakes location, genotypes like **WW6**, **WW1**, and **WW7** were observed to have the higher mean yields. Interestingly, the **WW7 genotype is in the top three highest-yielding groups among both the Aurora and Dakota Lakes** research farms (**Table 2**), indicating a promising level of **broad adaptation** across these two distinct environments.

Table 3: Descriptive statistics for grain yield (kg/ha) data of 10 wheat genotypes under four different nitrogen rates.

Location	N_Dose	Mean	SD	CV %	Minimum	Maximum	Median
Aurora	40lbs	2424.38	453.93	18.72	1365.30	3293.12	2440.79
Aurora	80lbs	2406.26	566.11	23.53	1169.54	3438.76	2564.56
Aurora	100lbs	2397.16	601.79	25.10	1041.68	3570.02	2359.45
Aurora	130lbs	2353.69	608.28	25.84	925.26	3108.74	2408.27
Dakota Lakes	40lbs	2527.75	410.13	16.22	1707.16	3647.41	2558.89
Dakota Lakes	80lbs	2491.15	279.55	11.22	1814.10	2963.21	2534.02
Dakota Lakes	100lbs	2592.47	312.38	12.05	1654.07	3082.55	2625.96
Dakota Lakes	130lbs	2453.13	352.75	14.38	1739.31	3072.34	2460.92

The yield response to Nitrogen (N) application was highly dependent on the location. At the Aurora Research Farm, the 40 lbs N level showed the highest mean yield (2424.38 kg/ha) and considerable variation (SD = 453.93) with a lower CV value (18.72 %). Crucially, there was a decreasing trend in mean yield as the N dose increased beyond 40 lbs/ac (the control). This finding is contrary to the expected positive response in N-limited environments and may indicate factors such as N toxicity, lodging, or nutrient imbalance interfering with yield at the higher application rates. Conversely, at the Dakota Lakes Research Farm, the 100 lbs dose showed the highest mean yield (2592.47 kg/ha) and the lowest variability (CV = 12.05 %) (**Table 3**). However, the overall N dose

was not highly effective at Dakota Lakes, as all doses showed similar, nearby mean yields, suggesting N was not the primary limiting factor at this site.

However, it is still necessary to utilize visual observation including the factors to have a better understanding. The first step of our analysis was to notice if the dependent variable shows some variation in the experiment.

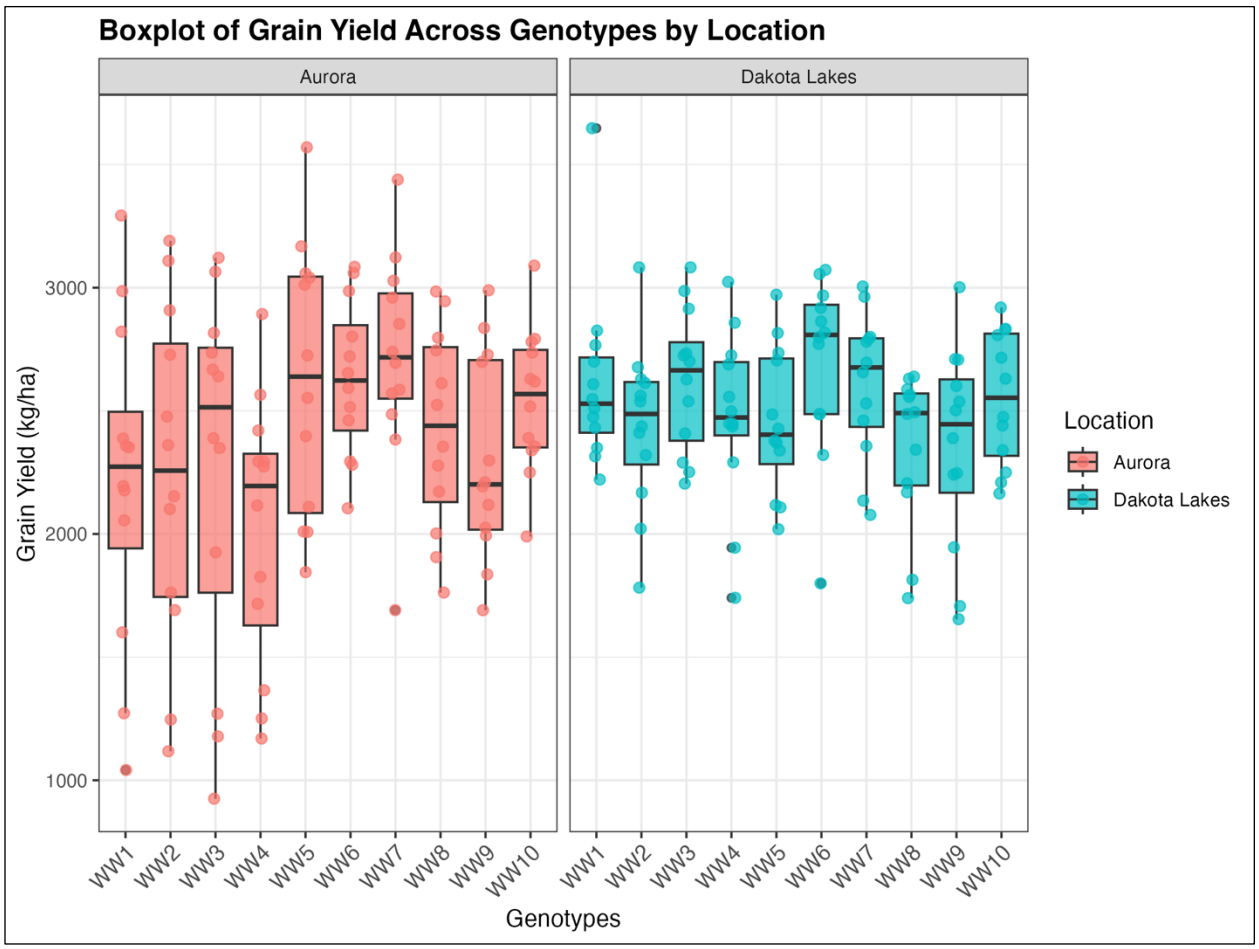


Fig.1: Boxplot of yield among evaluated 10 genotypes of two different locations.

The boxplot (**Fig. 1**) reveals the yield distribution for each genotype, highlighting the **Genotype x Location interaction**. At **Aurora**, **WW7** shows a high median yield with a relatively **narrower distribution** (indicating stability), followed by **WW6** and **WW5**. At **Dakota Lakes**, **WW7** again showed a high median yield, though its distribution was **wider** than some others, followed closely by **WW6** and **WW3**. Most importantly, WW7 was one of the **top median-yielding genotypes for both locations**, demonstrating **stability and broad adaptation** across the two environments.

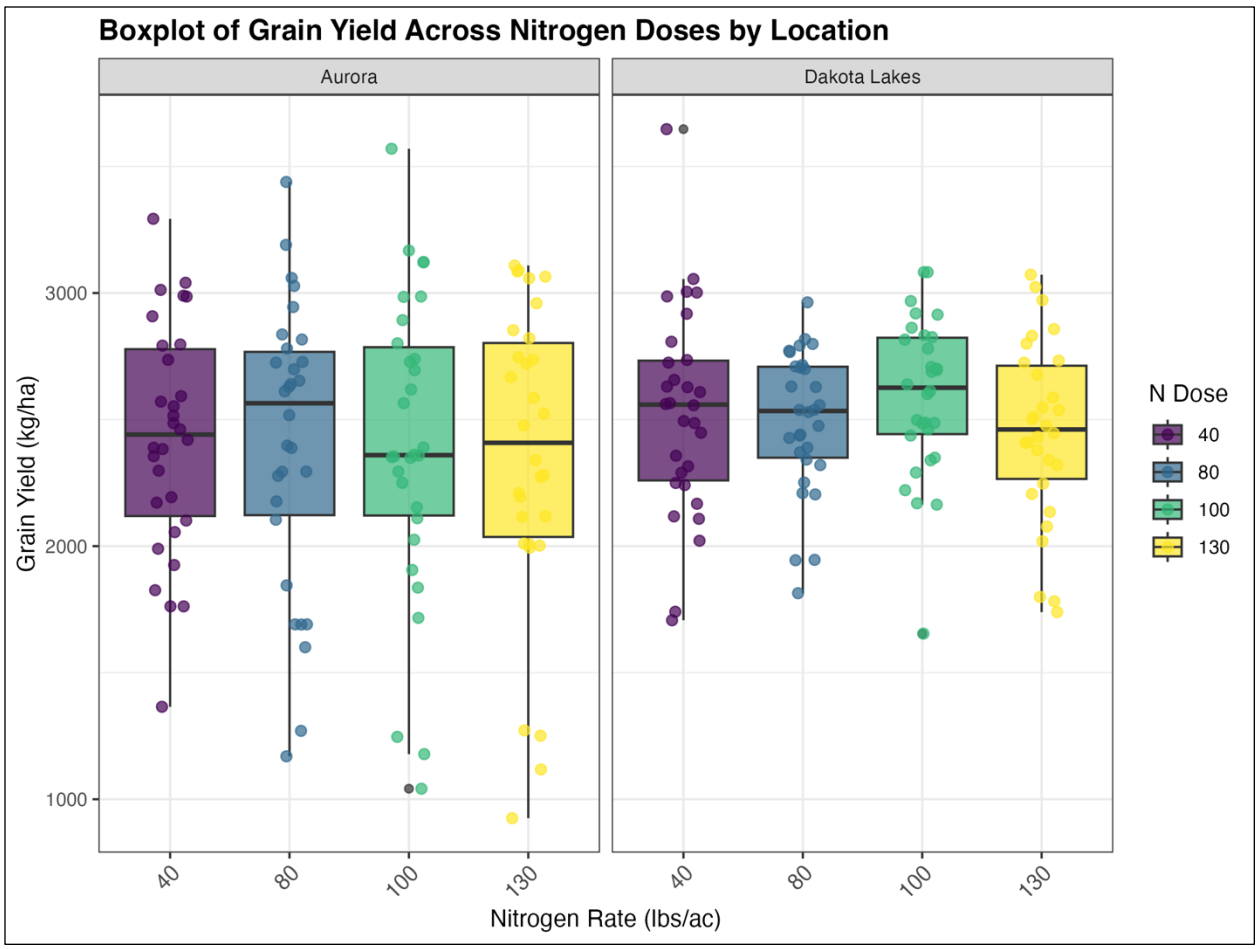


Fig.2: Boxplot of yield among four doses of Nitrogen at two different locations.

The boxplot (**Fig. 2**) reveals the yield distribution for each **Nitrogen Dose**. The median yield values at **Dakota Lakes** were consistently **higher** across all N doses and generally showed **narrower distributions** in comparison with the **Aurora** location. Overall, the **80 lbs/ac N Dose** showed the highest median yield at Aurora, while the **100 lbs/ac N Dose** showed the highest median yield at Dakota Lakes.

Our next step was to run an interaction plot (**Fig. 3**). At this time, it is possible to visualize in a better way how the factors could affect the dependent variable. The interaction plot shows how nitrogen levels affect yield across the genotypes (WW1–WW10) for two locations (Aurora and Dakota Lake). These locations exhibit the most variation in yield across nitrogen levels and genotypes. The lines crossing suggest a significant genotype × nitrogen interactions, meaning the response to nitrogen varies significantly across genotypes. This indicates that specific genotypes are more efficient at utilizing nitrogen.

The interaction plot shows that the lines connecting the N doses are not parallel and, critically, the pattern of these crossing lines **differs between the two locations**. For instance, at **Aurora**, several genotypes (e.g., WW3, WW2) show a sharp decline in yield when moving from 100 lbs/ac to 130 lbs/ac, possibly indicating a threshold or even N-toxicity effect in this specific environment. Conversely, at **Dakota Lakes**, such sharp declines are less frequent, and some genotypes (like WW1) respond strongly to higher N at this site.

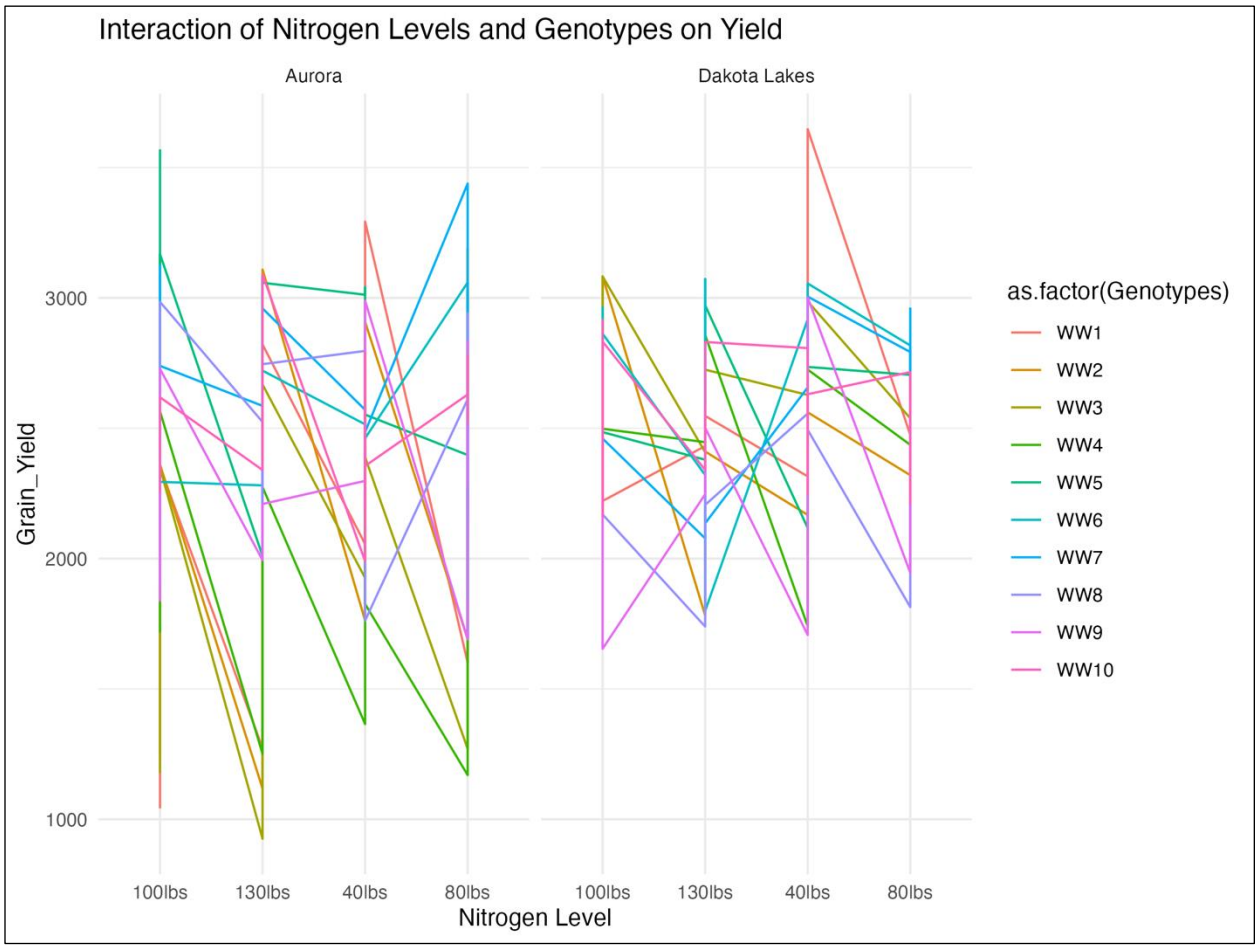
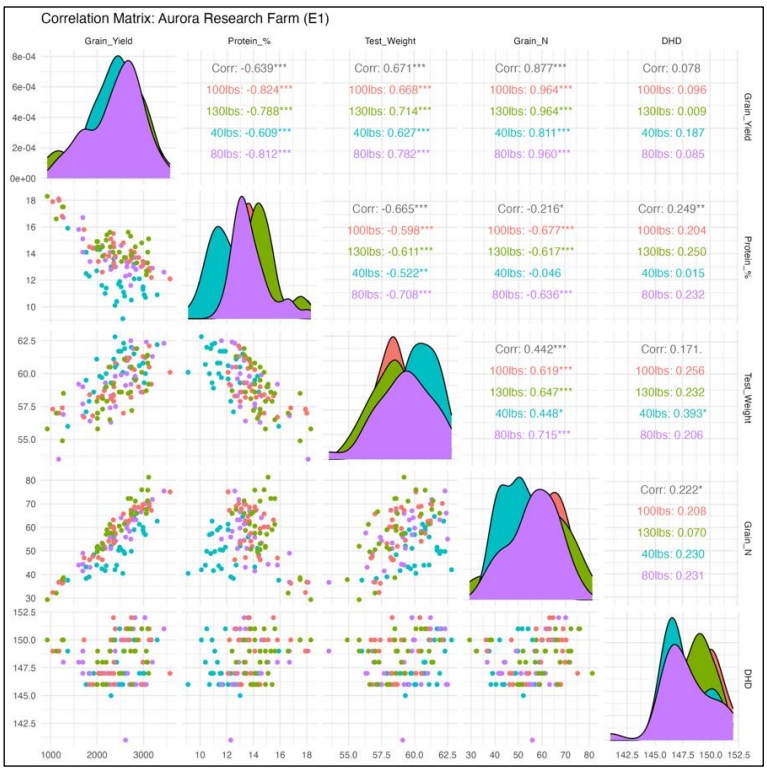


Fig.3: Interaction Plot of four different nitrogen doses (40lbs, 80 lbs, 100 lbs, 130 lbs) with 10 genotypes on yield at two different locations.

A)



B)

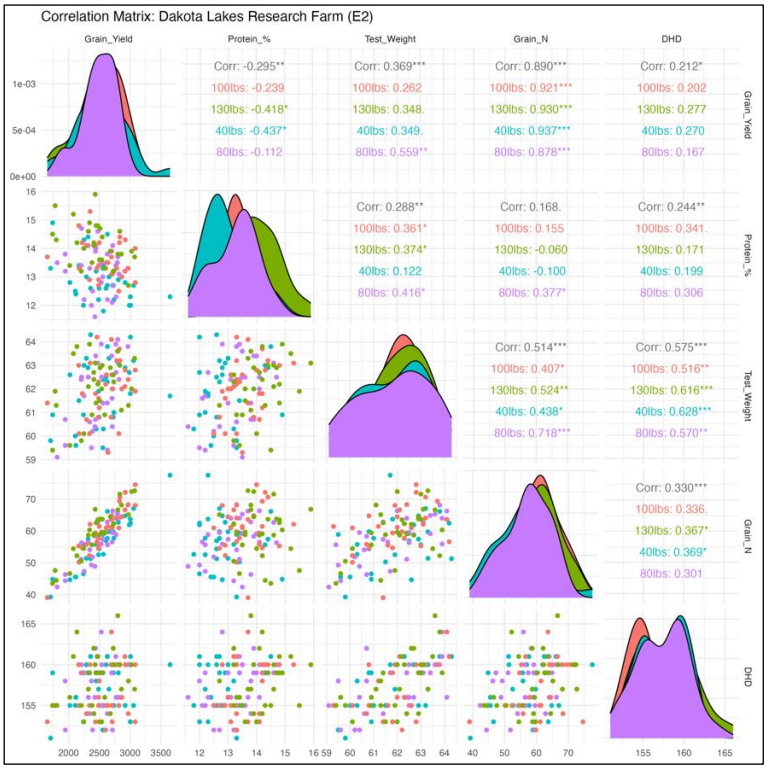


Fig. 4: The scatterplot matrix and correlation between the Yield and other traits (Protein, Grain Nitrogen (Grain_N), Test Weight, and Heading Date (DHD)) at different nitrogen application rates (40lbs, 80lbs, 100lbs, and 130lbs).

Yield vs. Protein: The correlation is consistently negative across all nitrogen levels for both the locations, with values ranging from -0.609 (40lbs) to -0.788 (130lbs) for Aurora and -0.437 (40lbs) to -0.418 (130lbs) for Dakota Lakes. This indicates that as protein content increases, yield tends to decrease. The relationship becomes stronger at higher nitrogen levels.

Yield vs. Grain_N: The correlation is very strong and positive across all nitrogen levels for both the locations, ranging from 0.811 (40lbs) to 0.964 (130lbs) for Aurora and 0.937 (40lbs) to 0.930 (130lbs) for Dakota Lakes. This suggests a significant association between grain nitrogen content and yield, meaning genotypes with higher grain nitrogen tend to yield more.

Yield vs. Test_Weight: A moderate positive correlation exists for both the locations, with values ranging from 0.627 (40lbs) to 0.714 (130lbs) for Aurora and 0.349 (40lbs) to 0.348 (130lbs) for Dakota Lakes. This indicates that higher test weight is associated with higher yield, though the strength of the relationship is weaker compared to Grain_N.

Yield vs. DHD (Heading Date): The correlation is weak and inconsistent across nitrogen levels, with values close to zero at both the locations. This suggests that heading date does not significantly influence yield in this dataset.

Protein vs. Grain_N: A moderate negative correlation exists, for both the locations. This indicates that as grain nitrogen increases, protein content decreases.

Protein vs. Test_Weight: A moderate negative correlation is observed for both the locations. Higher protein content is associated with lower test weight.

Grain_N vs. Test_Weight: A moderate positive correlation is observed for both the locations, ranging from 0.448 (40lbs) to 0.647 (130lbs) for Aurora and 0.438 (40lbs) to 0.524 (130lbs) for Dakota Lakes. This suggests that as grain nitrogen increases, test weight also tends to increase.

DHD (Heading Date) Relationships: Heading date shows weak or negligible correlations with all traits at both the locations, indicating that it has little influence on other parameters in this dataset.

Correlation patterns for most traits are consistent across nitrogen levels for both the locations, though the strength of relationships varies slightly. Higher nitrogen levels generally result in stronger correlations for traits like Yield vs. Grain_N and Yield vs. Protein, reflecting intensified

trait interactions at higher nutrient availability. To conclude anything in a statistics experiment, it is necessary to analyze with the right methodology. In our case, the right methodology was linear mixed model (LMM) for experimental design. Although we initially fitted a Generalized Linear Mixed Model (GLMM), diagnostic residual plots showed no evidence of overdispersion, non-normality, or heteroscedasticity that would necessitate a non-Gaussian distribution or link function. Moreover, the residuals from the GLMM were visually and statistically similar to those from a Linear Mixed Model (LMM), suggesting that the normality and homoscedasticity assumptions of the LMM were adequately met. Therefore, in accordance with the principle of parsimony, we opted for the simpler LMM, which provides equivalent fit with greater interpretability and computational efficiency.

4. Main statistical analysis (Model fitting, Inferences and Detailed Interpretations of Results)

4.1 Methodology Statement

We employed a Linear Mixed Model (LMM) using the lme4 package in R to evaluate the specific interactive effects of genetics and management practices on grain yield.

In this model, **Genotypes, Nitrogen Dose, and Location** were specified as **fixed effects** to enable direct mean comparisons and to assess the stability of specific genotypes across the varying nitrogen levels and the two distinct environments. Location was treated as fixed rather than random due to the limited number of test sites (n=2), which is insufficient for estimating environmental variance.

To account for the experimental design, **replications nested within locations** were treated as a **random effect**; this structure effectively partitioned the variance attributed to spatial field heterogeneity (blocking efficiency) from the residual error, thereby increasing the statistical power to detect significant differences among the fixed factors.

Model Equation

$$Y_{gdlr} = \mu + G_g + N_d + L_l + (G \times N)_{gd} + (G \times L)_{gl} + (N \times L)_{dl} + (G \times N \times L)_{gdl} + u_{lr} + \varepsilon_{gdlr}$$

Where:

- μ: overall mean
- G_g, N_d, L_l: main effects
- Interaction terms:
    - $(G \times N)_{gd}$,

- $(G \times L)_{gl}$,
- $(N \times L)_{dl}$,
- $(G \times N \times L)_{gdl}$

- $u_{lr} \sim N(0, \sigma_u^2)$: random effect of replication $r$within location l
- $\varepsilon_{gdlr} \sim N(0, \sigma^2)$: residual error

```
# Fit model
library(lme4)
model_yield <- lmer(
  Grain_Yield~ Genotypes * N_Dose * Location + (1 | Location:Replication), data = wheat_data)
```

4.2 Model Fitting and Assumptions:

To begin, we assessed the residuals of the linear mixed model using residual vs fitted plots and a Q-Q plot. The Q-Q plot showed a small deviation, indicating that the data were approximately normally distributed (**Fig. 5**). This confirmed that the linear mixed model (LMM) assumptions were satisfied.

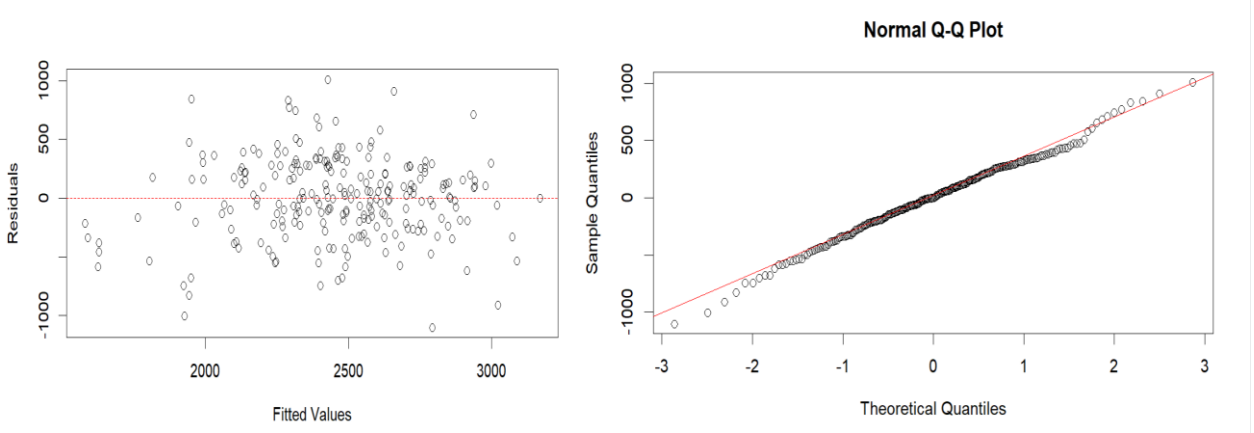


Fig. 5: Residual Vs fitted plot and Q-Q plot for checking model assumptions.

We can also test Shapiro-Wilk test, since our dataset has **240 observations** (10 Genotypes × 4 Doses × 2 Locations × 3 Reps), it fits perfectly within the limits of test (which works best for N < 5000). Checking this test is a also a valid way to confirm the normality assumption of the Linear Mixed Model residuals.

```
        Shapiro-Wilk normality test

data:  res
W = 0.99335, p-value = 0.3629
```

Residuals were checked for normality using the Shapiro-Wilk test **(W = 0.99, p = 0.36).** The test was non-significant, indicating that **the assumption of normality was met for the linear mixed model.** So, Model assumptions were verified through visual inspection of diagnostic plots. The plot of residuals versus fitted values exhibited a random scatter around the horizontal zero line with no discernible pattern, confirming the assumption of homoscedasticity (constant variance). Furthermore, the Normal Q-Q plot displayed residuals aligning closely with the diagonal reference line; this visual evidence, corroborated by a non-significant Shapiro-Wilk test (W = 0.99, p = 0.36), **confirms that the residuals follow a normal distribution, validating the use of the linear mixed model.**

4.3 Model Assessment: Multicollinearity

Next, we checked for multicollinearity using the Generalized Variance Inflation Factor (GVIF). Multicollinearity occurs when predictor variables are highly correlated, leading to instability in coefficient estimates.

Table 4: Generalized Variance Inflation factor (GVIF) for linear mixed model.

Variables	GVIF	DF	GVIF^(1/(2*Df))
Genotypes	1.342177e+08	9	2.828427
N_Dose	8.000000e+03	3	4.472136
Location	4.891005e+00	1	2.211562
Genotypes:N_Dose	1.143492e+16	27	1.983237
Genotypes:Location	1.944764e+08	9	2.887305
N_Dose:Location	9.197232e+03	3	4.577302
Genotypes:N_Dose:Location	1.261965e+16	27	1.986861

*GVIF refers to the Generalized Variance Inflation Factor, which measures multicollinearity; Df stands for Degrees of Freedom, indicating the number of independent values; and GVIF^(1/(2*Df)) is the adjusted GVIF, which normalizes the GVIF for degrees of freedom.*

So, we observed moderate collinearity for most of the predictors with GVIF values close to 2 and 4, respectively (**Table 4**). These values are generally acceptable, though we noted that the interaction term between **Genotypes:N_Dose:Location** had a high GVIF (1.261965e+16), which might be problematic. Upon further inspection, we decided to remove this interaction term as well, as it was causing multicollinearity and affecting model stability.

```
model_Simplified <- lmer(
  Grain_Yield ~ Genotypes * Location + Genotypes * N_Dose + N_Dose * Location +
    (1 | Location:Replication),
  data = wheat_data
)
```

Model equation

$$Y_{gdlr} = \mu + G_g + N_d + L_l + (G \times L)_{gl} + (G \times N)_{gd} + (N \times L)_{dl} + u_{lr} + \varepsilon_{gdlr}$$

Where:

- μ: overall mean
- G_g, N_d, L_l: main effects
- $(G \times L)_{gl}, (G \times N)_{gd}, (N \times L)_{dl}$: two-way interactions
- $u_{lr} \sim N(0, \sigma_u^2)$: random replication effect nested within location
- $\varepsilon_{gdlr} \sim N(0, \sigma^2)$: residual error

Table 5: Generalized Variance Inflation factor (GVIF) for linear mixed model.

Variables	GVIF	DF	GVIF^(1/(2*Df))
Genotypes	1.953125e+06	9	2.236068
N_Dose	2.150607e+00	3	4.472136
Location	1.331000e+03	1	2.211562
Genotypes:N_Dose	7.329165e+02	27	1.983237
Genotypes:Location	8.519680e+07	9	2.887305
N_Dose:Location	9.150607e+00	3	4.577302

*GVIF refers to the Generalized Variance Inflation Factor, which measures multicollinearity; Df stands for Degrees of Freedom, indicating the number of independent values; and GVIF^(1/(2*Df)) is the adjusted GVIF, which normalizes the GVIF for degrees of freedom.*

After removing **Genotypes:N_Dose:Location,** the variance inflation factor (VIF) analysis for the final model indicates **low to moderate multicollinearity** among predictors (**Table 5**). Most of the terms exhibit **GVIF^(1/(2·Df)) values around 2**, suggesting minimal collinearity, which is still within acceptable limits for controlled agronomic trials and unlikely to compromise inference. Overall, the model demonstrates **sufficient numerical stability**, and parameter estimates can be interpreted with reasonable confidence.

To evaluate the adequacy of this multi-trait model, we again performed residual analysis (**Fig. 6**). The residual plot for the simplified linear mixed model showed a more uniform spread, indicating that the model's assumptions were better satisfied compared to the initial linear mixed model. The spread of the residuals was more even, with no obvious patterns, suggesting that the model appropriately captured the relationships in the data. This improvement in the residual plot further supports the choice of the simplified linear mixed model.

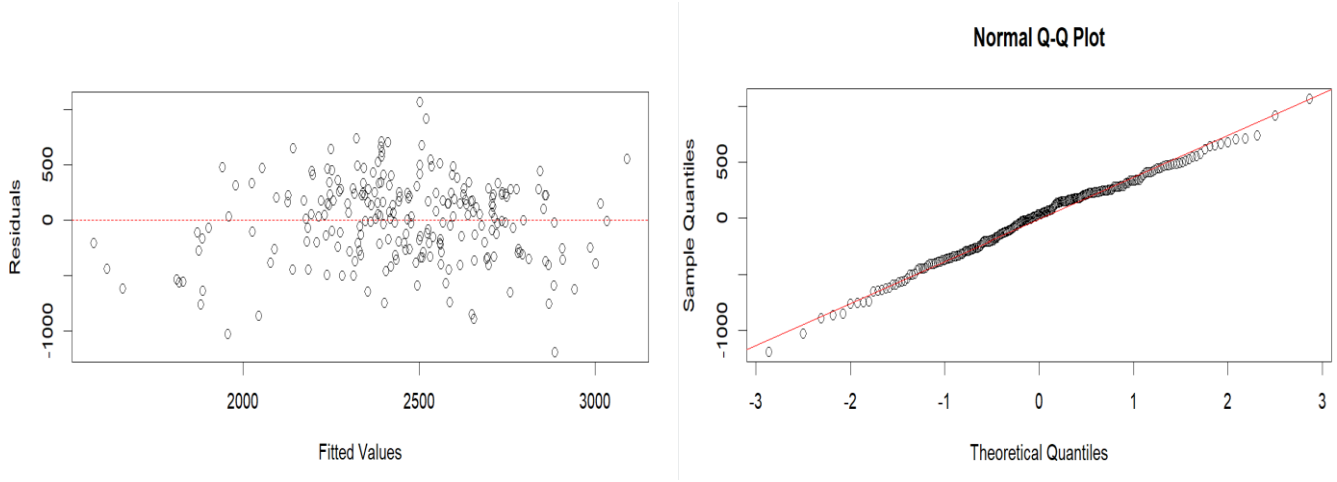


Fig. 6: Residual Vs fitted plot and Q-Q plot for checking model assumptions.

4.4 Post-Hoc Comparisons: Genotype and Nitrogen Rate Effects

To further explore the differences among genotypes and nitrogen doses, location and with their interactions, we first checked anova table (**Table 6**):

```
> anova(model_Simplified)
Type III Analysis of Variance Table with Satterthwaite's method
                    Sum Sq Mean Sq NumDF DenDF F value  Pr(>F)
Genotypes          4362564  484729     9   183  2.7338 0.00513 **
Location             83890   83890     1     4  0.4731 0.52937
N_Dose              283651   94550     3   183  0.5332 0.66004
Genotypes:Location 2710349  301150     9   183  1.6984 0.09209 .
Genotypes:N_Dose   3166801  117289    27   183  0.6615 0.89786
Location:N_Dose     114017   38006     3   183  0.2143 0.88637
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 6: Analysis of variance of the simplified model

From the anova table we came to know that there is significant difference between the genotypes:

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
WW1 - WW10	-131.97	122	183	-521.3	257.4	-1.086	0.9856
WW1 - WW2	77.25	122	183	-312.1	466.6	0.636	0.9998
WW1 - WW3	-25.35	122	183	-414.7	364.0	-0.209	1.0000
WW1 - WW4	170.71	122	183	-218.7	560.1	1.404	0.9246
WW1 - WW5	-126.40	122	183	-515.8	263.0	-1.040	0.9894
WW1 - WW6	-249.01	122	183	-638.4	140.4	-2.049	0.5661
WW1 - WW7	-244.61	122	183	-634.0	144.8	-2.012	0.5916
WW1 - WW8	26.04	122	183	-363.3	415.4	0.214	1.0000
WW1 - WW9	86.52	122	183	-302.8	475.9	0.712	0.9994
WW10 - WW2	209.22	122	183	-180.1	598.6	1.721	0.7821
WW10 - WW3	106.63	122	183	-282.7	496.0	0.877	0.9970
WW10 - WW4	302.68	122	183	-86.7	692.0	2.490	0.2809
WW10 - WW5	5.57	122	183	-383.8	394.9	0.046	1.0000
WW10 - WW6	-117.04	122	183	-506.4	272.3	-0.963	0.9939
WW10 - WW7	-112.64	122	183	-502.0	276.7	-0.927	0.9954
WW10 - WW8	158.01	122	183	-231.4	547.4	1.300	0.9526
WW10 - WW9	218.49	122	183	-170.9	607.9	1.797	0.7361
WW2 - WW3	-102.60	122	183	-492.0	286.8	-0.844	0.9977
WW2 - WW4	93.46	122	183	-295.9	482.8	0.769	0.9989
WW2 - WW5	-203.65	122	183	-593.0	185.7	-1.675	0.8079
WW2 - WW6	-326.26	122	183	-715.6	63.1	-2.684	0.1881
WW2 - WW7	-321.86	122	183	-711.2	67.5	-2.648	0.2035
WW2 - WW8	-51.21	122	183	-440.6	338.2	-0.421	1.0000
WW2 - WW9	9.27	122	183	-380.1	398.6	0.076	1.0000
WW3 - WW4	196.05	122	183	-193.3	585.4	1.613	0.8404
WW3 - WW5	-101.05	122	183	-490.4	288.3	-0.831	0.9980
WW3 - WW6	-223.66	122	183	-613.0	165.7	-1.840	0.7089
WW3 - WW7	-219.26	122	183	-608.6	170.1	-1.804	0.7321
WW3 - WW8	51.39	122	183	-338.0	440.8	0.423	1.0000
WW3 - WW9	111.86	122	183	-277.5	501.2	0.920	0.9956
WW4 - WW5	-297.11	122	183	-686.5	92.3	-2.444	0.3063
WW4 - WW6	-419.72	122	183	-809.1	-30.4	-3.453	0.0235
WW4 - WW7	-415.32	122	183	-804.7	-25.9	-3.417	0.0263
WW4 - WW8	-144.67	122	183	-534.0	244.7	-1.190	0.9731
WW4 - WW9	-84.19	122	183	-473.6	305.2	-0.693	0.9995
WW5 - WW6	-122.61	122	183	-512.0	266.8	-1.009	0.9914
WW5 - WW7	-118.21	122	183	-507.6	271.2	-0.972	0.9934
WW5 - WW8	152.44	122	183	-236.9	541.8	1.254	0.9622
WW5 - WW9	212.92	122	183	-176.4	602.3	1.752	0.7642
WW6 - WW7	4.40	122	183	-385.0	393.8	0.036	1.0000
WW6 - WW8	275.05	122	183	-114.3	664.4	2.263	0.4184
WW6 - WW9	335.53	122	183	-53.8	724.9	2.760	0.1583
WW7 - WW8	270.65	122	183	-118.7	660.0	2.227	0.4426
WW7 - WW9	331.13	122	183	-58.2	720.5	2.724	0.1720
WW8 - WW9	60.48	122	183	-328.9	449.8	0.498	1.0000

We conducted post-hoc analyses using the best-fitting simplified linear mixed model. Significant differences were found between certain genotype pairs, as summarized below:

	contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
32	WW4 - WW6	-419.7191	121.5567	183	-809.0869	-30.35124	-3.452867	0.02351159
33	WW4 - WW7	-415.3153	121.5567	183	-804.6831	-25.94744	-3.416639	0.02634605

5. Summary of Recommendations:

The integrated analysis reveals that genotype is the predominant factor driving grain yield variation, with **WW7 emerging as a consistently high-performing and broadly adapted line across both Aurora and Dakota Lakes locations**: supported by high mean yields, low yield variability (CV), and top median yields in boxplots. While nitrogen (N) application effects were location-dependent: showing

an unexpected yield decline beyond 40 lbs N/ac at Aurora but a modest peak at 100 lbs N/ac at Dakota Lakes**; overall N dose and location lacked significant main effects in the ANOVA, and their interactions with genotype were non-significant, suggesting stable genotype rankings.** However, interaction plots revealed non-parallel response patterns across genotypes and N levels, hinting at genotype-specific N use efficiency that warrants deeper investigation despite non-significant statistical interactions. Trait correlations further clarify physiological trade-offs: **a strong positive link between grain N and yield**, a **consistent negative yield–protein relationship** (intensifying with higher N), and **moderate positive associations between yield and test weight**, while **heading date showed negligible influence**. ANOVA identified genotype as the sole significant factor affecting grain yield ($F = 2.73$, $p = 0.0051$), with no significant main effects of location ($p = 0.529$) or nitrogen dose ($p = 0.660$), nor any significant interactions (all $p > 0.09$).This indicates consistent genotype performance across the two locations (Aurora, Dakota Lakes) and three N rates, reflecting high stability. Post-hoc tests revealed **WW6 and WW7 significantly outperformed the control WW4** ($p < 0.05$), while other genotypes showed similar yields. Results confirm that genetic selection, not environment or N management: drives yield variation, supporting **the broad adaptability of top lines like WW6 and WW7.** Overall, these findings emphasize that **genetic selection remains the primary driver of yield variation** in this trial, and that **environmental factors (location, N dose) and their interactions did not substantially modify genotype ranking**: supporting the potential for deploying top-performing genotypes broadly without site- or dose-specific adjustments.

Conclusion:

These findings demonstrate that **genetic selection is the dominant lever for improving wheat grain yield** under the tested conditions, with WW6 and especially WW7 exhibiting **high yield, stability, and broad adaptability** across contrasting environments and N management strategies. The lack of significant genotype × environment interactions supports the **feasibility of deploying these elite lines widely without location- or N-specific adjustments**. However, visual trends in interaction plots and location-specific N responses; despite statistical non-significance - suggest underlying physiological differences in N use efficiency that merit targeted investigation. Future work should prioritize validating WW6 and WW7 in expanded environments, dissecting their N-response physiology, and addressing the yield–protein trade-off to develop varieties that combine high productivity with end-use quality. For now, these genotypes represent promising candidates for breeding programs and on-farm adoption in similar production systems.

6. References:

Zhang, Zhen, et al. "Optimized nitrogen fertilizer application strategies under supplementary irrigation improved winter wheat (Triticum aestivum L.) yield and grain protein yield." PeerJ 9 (2021): e11467.

Cassman, K. G., Dobermann, A., & Walters, D. T. (2002). Agroecosystems, nitrogen-use efficiency, and nitrogen management. AMBIO: A Journal of the Human Environment, 31(2), 132-140.

Parent, B., Bonneau, J., Maphosa, L., Kovalchuk, A., Langridge, P., & Fleury, D. (2017). Quantifying wheat sensitivities to environmental constraints to dissect genotype× environment interactions in the field. Plant Physiology, 174(3), 1669-1682.

7. Appendix

R script used for the above data analysis is given below:

```
setwd("~/Desktop")
list.files()
df <- read.xlsx("AAHS_Data  (1).xlsx")
getSheetNames("AAHS_Data  (1).xlsx")
df <- read.xlsx("AAHS_Data  (1).xlsx","Winter_Wheat_AAHS_Data")
head(df)
##Descriptive statistics of genotypes
library(dplyr)
summary_grouped <- df %>%
  group_by(Location, Year, Genotypes) %>%   # fix name if needed
  summarise(across(where(is.numeric),
                   list(
                     mean = mean,
                     sd = sd,
                     min = min,
                     max = max,
                     median = median
                   ),
                   .names = "{.col}_{.fn}"),
            .groups = "drop")

summary_grouped
print(summary_grouped)
library(writexl)
write_xlsx(summary_grouped, "summary_grouped.xlsx")

#################
#Descriptive statistics of Nitrogen dose
summary_grouped <- df %>%
  group_by(Location, Year, N_Dose) %>%   # fix name if needed
  summarise(across(where(is.numeric),
                   list(
                     mean = mean,
                     sd = sd,
                     min = min,
```

```
                max = max,
                median = median
              ),
              .names = "{.col}_{.fn}"),
          .groups = "drop")
summary_grouped
print(summary_grouped)
write_xlsx(summary_grouped, "N_Dose_summary_grouped.xlsx")

############

# Outliers Function (IQR method)
detect_outliers <- function(x) {
  q1 <- quantile(x, 0.25, na.rm = TRUE)
  q3 <- quantile(x, 0.75, na.rm = TRUE)
  iqr <- q3 - q1
  lower <- q1 - 1.5 * iqr
  upper <- q3 + 1.5 * iqr
  return(x < lower | x > upper)
}
# Remove missing values first
data <- df %>%
  drop_na(Grain_Yield)
# Detect outliers (correct grouping)
data_outliers <- data %>%
  group_by(Genotypes) %>%
  mutate(is_outlier = detect_outliers(Grain_Yield))
print(data_outliers)

################################
#Boxplot of Grain Yield Across Genotypes by Location
library(ggplot2)
library(dplyr)
# Ensure the Genotypes are ordered WW1 through WW10
genotype_order <- paste0("WW", 1:10)
df$Genotypes <- factor(df$Genotypes, levels = genotype_order)
# Boxplot Plotting Code with Jitter and FIXED SCALES ---
final_genotypes_plot <- ggplot(df, aes(x = Genotypes, y = Grain_Yield)) +
  # Use different colors/fills for the boxes to distinguish the environments
  geom_boxplot(aes(fill = Location), alpha = 0.7) +
    # Geom Jitter for individual data points
  geom_jitter(
    aes(color = Location),
    size = 2,
    width = 0.1,
    alpha = 0.7
  ) +
    # Separate the plot into two panels, one for each location
  facet_wrap(~ Location, scales = "fixed") +
    labs(
    title = "Boxplot of Grain Yield Across Genotypes by Location",
```

```
    x = "Genotypes",
    y = "Grain Yield (kg/ha)",
    fill = "Location",
    color = "Location"
  ) +
  theme_bw() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 10),
    plot.title = element_text(face = "bold")
  )


ggsave("Figure_3_3_Yield_by_Geno_Location.png",
       plot = final_genotypes_plot,
       width = 8,
       height = 6,
       dpi = 300)
print(final_genotypes_plot)

#################
#Boxplot of Grain Yield Across Nitrogen Doses by Location
library(ggplot2)
library(dplyr)
# --- Data Cleaning and Explicit Sequencing FIX ---
# 1. CLEANING: Assuming N_Dose still contains "lbs" (e.g., "40lbs"), we must clean it first.
#    If the column is named 'N_Dose/ac' in your raw data, adjust the line below.
df$N_Dose_clean <- as.numeric(gsub("lbs", "", df$N_Dose))
# 2. SEQUENCING FIX: Define the order based on the cleaned numeric values.
n_dose_order <- c(40, 80, 100, 130)
df$N_Dose_f <- factor(df$N_Dose_clean, levels = n_dose_order, ordered = TRUE)
# --- Plotting Code (using the new N_Dose_f factor and correct Y-axis name) ---
final_ndose_plot <- ggplot(df, aes(x = N_Dose_f, y = Grain_Yield)) +
    geom_boxplot(aes(fill = N_Dose_f), alpha = 0.7) +
    geom_jitter(
    aes(color = N_Dose_f),
    width = 0.15,
    size = 2,
    alpha = 0.7
  ) +
    facet_wrap(~ Location, scales = "fixed") +
    labs(
    title = "Boxplot of Grain Yield Across Nitrogen Doses by Location",
    x = "Nitrogen Rate (lbs/ac)",
    y = "Grain Yield (kg/ha)",
    fill = "N Dose",
    color = "N Dose"
  ) +
    theme_bw() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 10),
    plot.title = element_text(face = "bold")
```

```
  )
ggsave("Figure_3_3_Yield_by_NDose_Location.png",
       plot = final_ndose_plot,
       width = 8,
       height = 6,
       dpi = 300)
print(final_ndose_plot)

################################
#interaction plots
#intraction of nitrogen levels and Genetypes
final_interaction_plot <- ggplot(df, aes(x = N_Dose, y = Grain_Yield, color = as.factor(Genotypes))) +
  geom_line(aes(group = Genotypes)) +
  facet_wrap(~Location) +
  labs(title = "Interaction of Nitrogen Levels and Genotypes on Yield",
       x = "Nitrogen Level", y = "Grain_Yield") +
  theme_minimal()
 ggsave("Figure_interaction_NDose_Location.png",
       plot = final_interaction_plot,
       width = 8,
       height = 6,
       dpi = 300)
print(final_interaction_plot)

#####################################
#Correlation matrix
library(GGally)
library(dplyr)
library(ggplot2) # Needed for ggsave
# Subset data for Aurora
df_aurora <- df %>% filter(Location == "Aurora")
# Subset data for Dakota Lakes
df_dakota <- df %>% filter(Location == "Dakota Lakes")
# Define the columns (Yield, Protein, Test_Weight, Grain_N, Days_to_Heading)
# Based on the report structure:
var_cols <- c("Grain_Yield", "Protein_%", "Test_Weight",
              "Grain_N", "DHD")
# AURORA CORRELATION MATRIX ---
plot_aurora <- ggpairs(
  df_aurora,
  columns = var_cols,
  mapping = aes(color = N_Dose), # Grouping by the N-Dose factor for color
  title = "Correlation Matrix: Aurora Research Farm (E1)"
) +
  theme_minimal()
# DAKOTA LAKES CORRELATION MATRIX ---
plot_dakota <- ggpairs(
  df_dakota,
  columns = var_cols,
  mapping = aes(color = N_Dose), # Grouping by the N-Dose factor for color
  title = "Correlation Matrix: Dakota Lakes Research Farm (E2)"
```

```
) +
  theme_minimal()
# DISPLAY AND SAVE ---
# Display the plots (you will need to display them sequentially in RStudio)
print(plot_aurora)
print(plot_dakota)
# Save plots as separate PNG files for the Appendix
ggsave("Appendix_Figure_Correlation_Aurora.png", plot = plot_aurora, width = 10, height = 10, dpi =
300)
ggsave("Appendix_Figure_Correlation_DakotaLakes.png", plot = plot_dakota, width = 10, height = 10,
dpi = 300)
wheat_data <- read_excel("Book1.xlsx")
# Convert to factors
wheat_data <- read_excel("Book1.xlsx")
wheat_data$Genotypes <- as.factor(wheat_data$Genotypes)
wheat_data$N_Dose <- as.factor(wheat_data$N_Dose)
wheat_data$Location <- as.factor(wheat_data$Location)
wheat_data$Replication <- as.factor(wheat_data$Replication)
# Fit model
library(lme4)
model_yield <- lmer(
  Grain_Yield~ Genotypes * N_Dose * Location + (1 | Location:Replication), data = wheat_data)
# Summary
summary(model_yield)
anova(model_yield)
# After fitting your model (e.g., model_yield)
res <- residuals(model_yield)
fitted_vals <- fitted(model_yield)
# Residuals vs Fitted
plot(fitted_vals, res, xlab = "Fitted Values", ylab = "Residuals")
abline(h = 0, col = "red", lty = 2)
# Q-Q Plot
qqnorm(res)
qqline(res, col = "red")  # Better than abline(0,1) for residuals
# Calculate residuals (if you haven't already)
res <- residuals(model_yield)
# Run the Shapiro-Wilk test
shapiro.test(res)
library(car)
vif(model_yield)
model_Simplified <- lmer(
  Grain_Yield ~ Genotypes * Location + Genotypes * N_Dose + N_Dose * Location +
    (1 | Location:Replication),
  data = wheat_data
)
vif(model_Simplified)
anova(model_Simplified)
library(emmeans)
res <- residuals(model_Simplified)
fitted_vals <- fitted(model_Simplified)
# Residuals vs Fitted
```

```
plot(fitted_vals, res, xlab = "Fitted Values", ylab = "Residuals")
abline(h = 0, col = "red", lty = 2)
# Q-Q Plot
qqnorm(res)
qqline(res, col = "red")  # Better than abline(0,1) for residuals
# Compare genotypes averaged over N_Dose and Location
emm_gen <- emmeans(model_Simplified, ~ Genotypes)
pairs_gen <- pairs(emm_gen, adjust = "tukey")
# View results
summary(pairs_gen, infer = TRUE)
# Get full summary with CIs and p-values
posthoc_table <- summary(pairs_gen, infer = TRUE)
# Extract significant ones
sig_only <- subset(posthoc_table, p.value < 0.05)
print(sig_only)
summary(model_yield)
anova(model_yield)
# After fitting your model (e.g., model_yield)
res <- residuals(mixed_model_3way)
fitted_vals <- fitted(mixed_model_3way)
# Residuals vs Fitted
plot(fitted_vals, res, xlab = "Fitted Values", ylab = "Residuals")
abline(h = 0, col = "red", lty = 2)
# Q-Q Plot
qqnorm(res)
qqline(res, col = "red")  # Better than abline(0,1) for residuals
library(car)
vif(mixed_model_3way)
anova(mixed_model_3way)
library(emmeans)
# Compare genotypes averaged over N_Dose and Location
emm_gen <- emmeans(model_Simplified, ~ Genotypes)
pairs_gen <- pairs(emm_gen, adjust = "tukey")
# View results
summary(pairs_gen, infer = TRUE)
# Get full summary with CIs and p-values
posthoc_table <- summary(pairs_gen, infer = TRUE)
# Extract significant ones
sig_only <- subset(posthoc_table, p.value < 0.05)
print(sig_only)
```