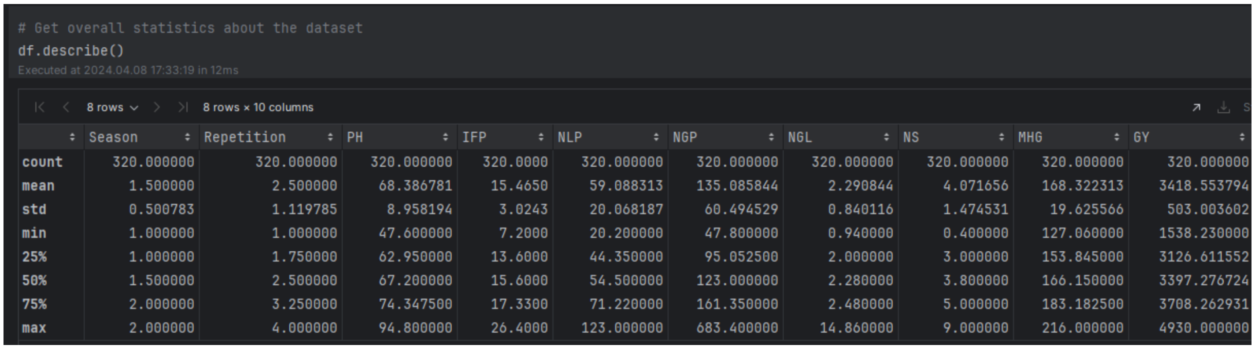
Table of contents

1. Introduction.

This documentation presents an overview of a machine learning project focused on soybean cultivation. The project aims to utilize data-driven approaches to predict the thousand seed weight of soybeans and to cluster similar cultivars.

2. EDA.

2.1. Basic info. The dataset comprises 11 columns and 320 entries, predominantly consisting of floating-point values, with exceptions including "Repetition," "Season," and "Cultivar." Notably, the mean values of the thousand seed weight (MHG) and grain yield (GY) within the dataset are calculated to be ~168 and ~3418.

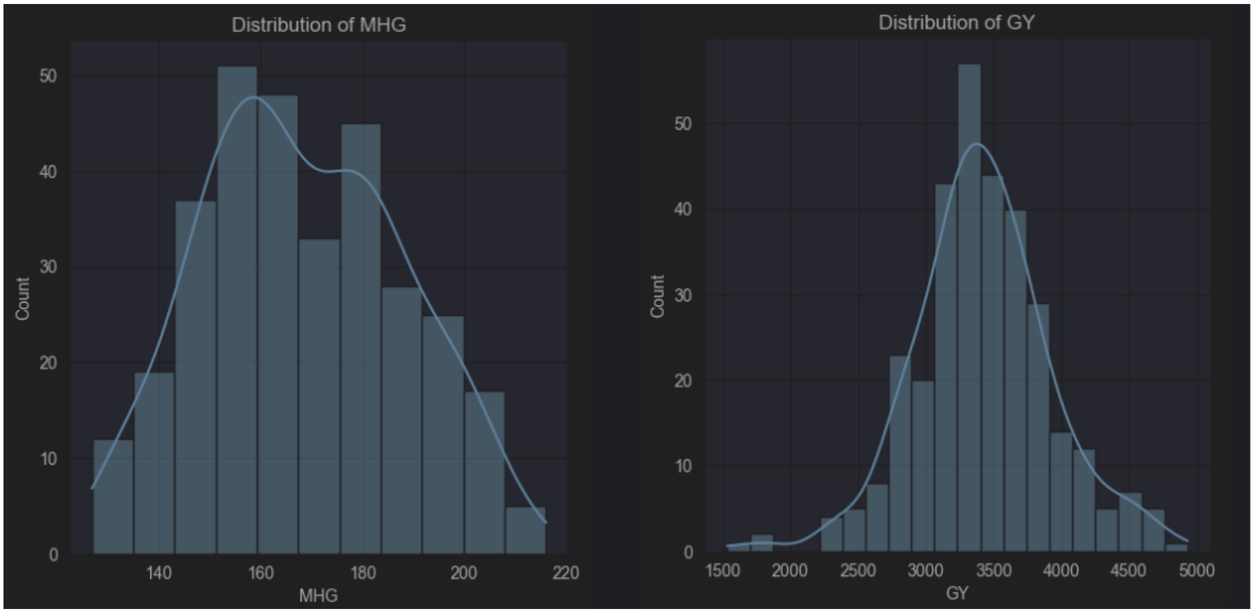


```
# Get overall statistics about the dataset
df.describe()
```

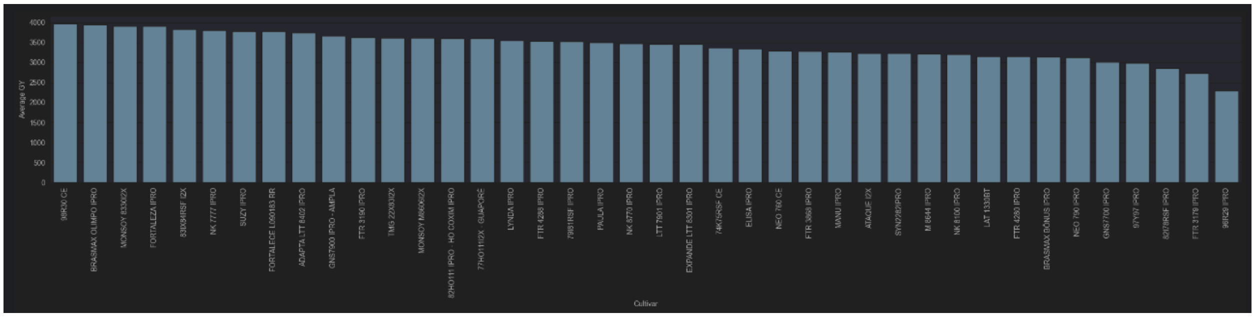
Executed at 2024.04.08 17:33:19 in 12ms

	Season	Repetition	PH	IFP	NLP	NGP	NGL	NS	MHG	GY
count	320.000000	320.000000	320.000000	320.0000	320.000000	320.000000	320.000000	320.000000	320.000000	320.000000
mean	1.500000	2.500000	68.386781	15.4650	59.088313	135.085844	2.290844	4.071656	168.322313	3418.553794
std	0.500783	1.119785	8.958194	3.0243	20.068187	60.494529	0.840116	1.474531	19.625566	503.003602
min	1.000000	1.000000	47.600000	7.2000	20.200000	47.800000	0.940000	0.400000	127.060000	1538.230000
25%	1.000000	1.750000	62.950000	13.6000	44.350000	95.052500	2.000000	3.000000	153.845000	3126.611552
50%	1.500000	2.500000	67.200000	15.6000	54.500000	123.000000	2.280000	3.800000	166.150000	3397.276724
75%	2.000000	3.250000	74.347500	17.3300	71.220000	161.350000	2.480000	5.000000	183.182500	3708.262931
max	2.000000	4.000000	94.800000	26.4000	123.000000	683.400000	14.860000	9.000000	216.000000	4930.000000

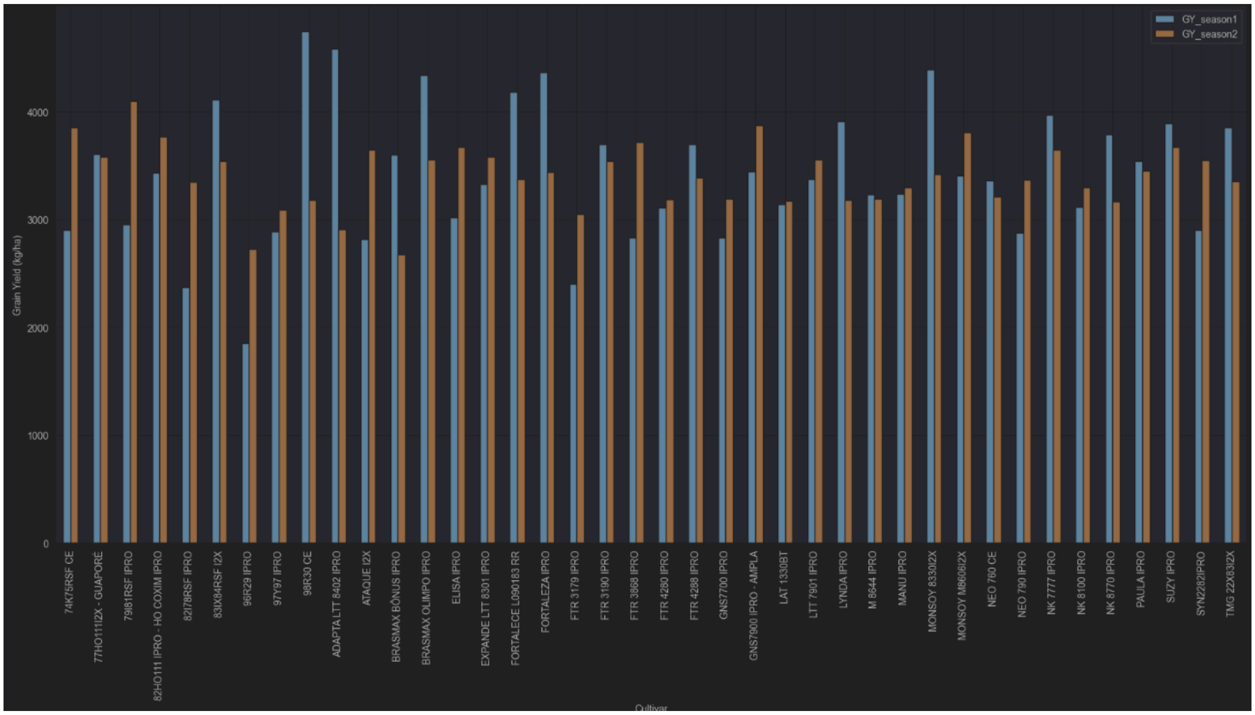
2.2. Data distribution. Next, we provide histograms depicting the distribution of data for each variable, to visualize the frequency distribution of values which provides insights itno the data's spread and central tendencies. For example, here are the histograms for MHG and GY:



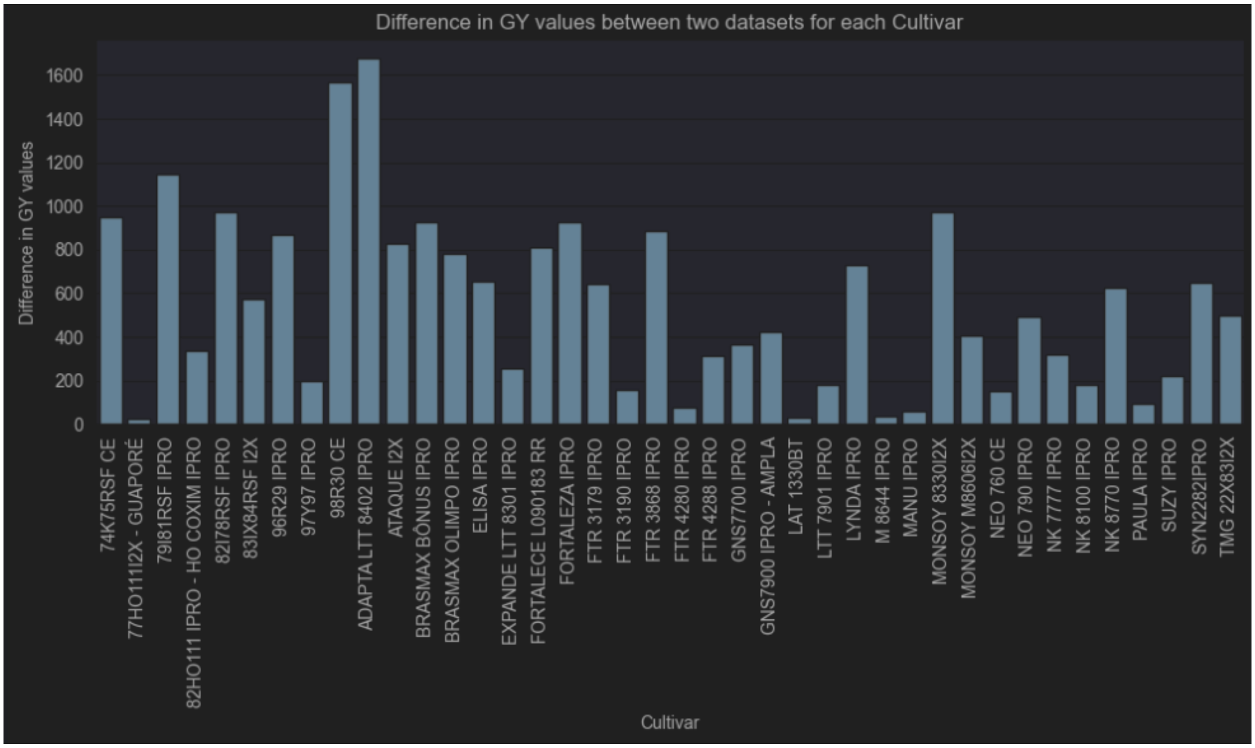
2.3. Top cultivars by grain yield. In this section, an analysis of cultivar performance is presented through a bar chart. The horizontal axis represents different cultivars, while the vertical axis represents grain yield (GY), computed as the average GY produced by each cultivar over two seasons, with four repetitions for each season. The purpose is to provide a succint overview of cultivar performance in terms of grain yield. By aggregating data across multiple seasons and repetitions, this analysis aims to identify the cultivars exhibiting the highest average grain yield.



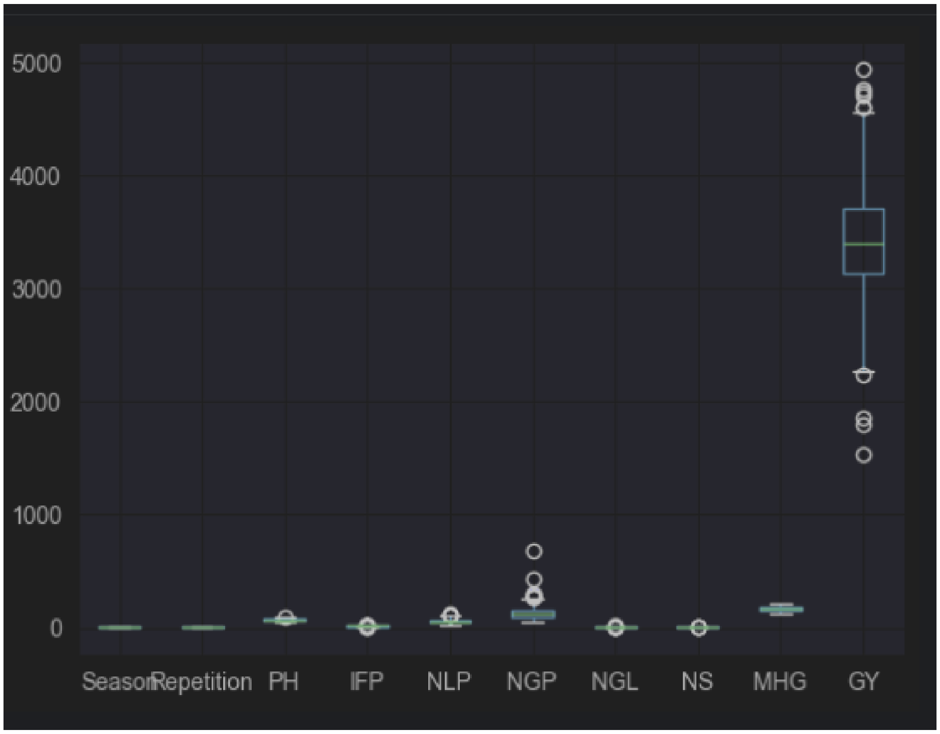
2.4. Difference in grain yield (GY) and thousand seed weight (MHG) in season 1 and season 2. In this section, we examine the difference between GY and MHG across the two seasons. Two distinct bar charts are presented. The first bar chart showcases the GY produced by each cultivar in the two seasons which enables us to visualize a comparative analysis of GY across seasons.



The second bar chart shows us the absolute value of the difference between the two seasons.



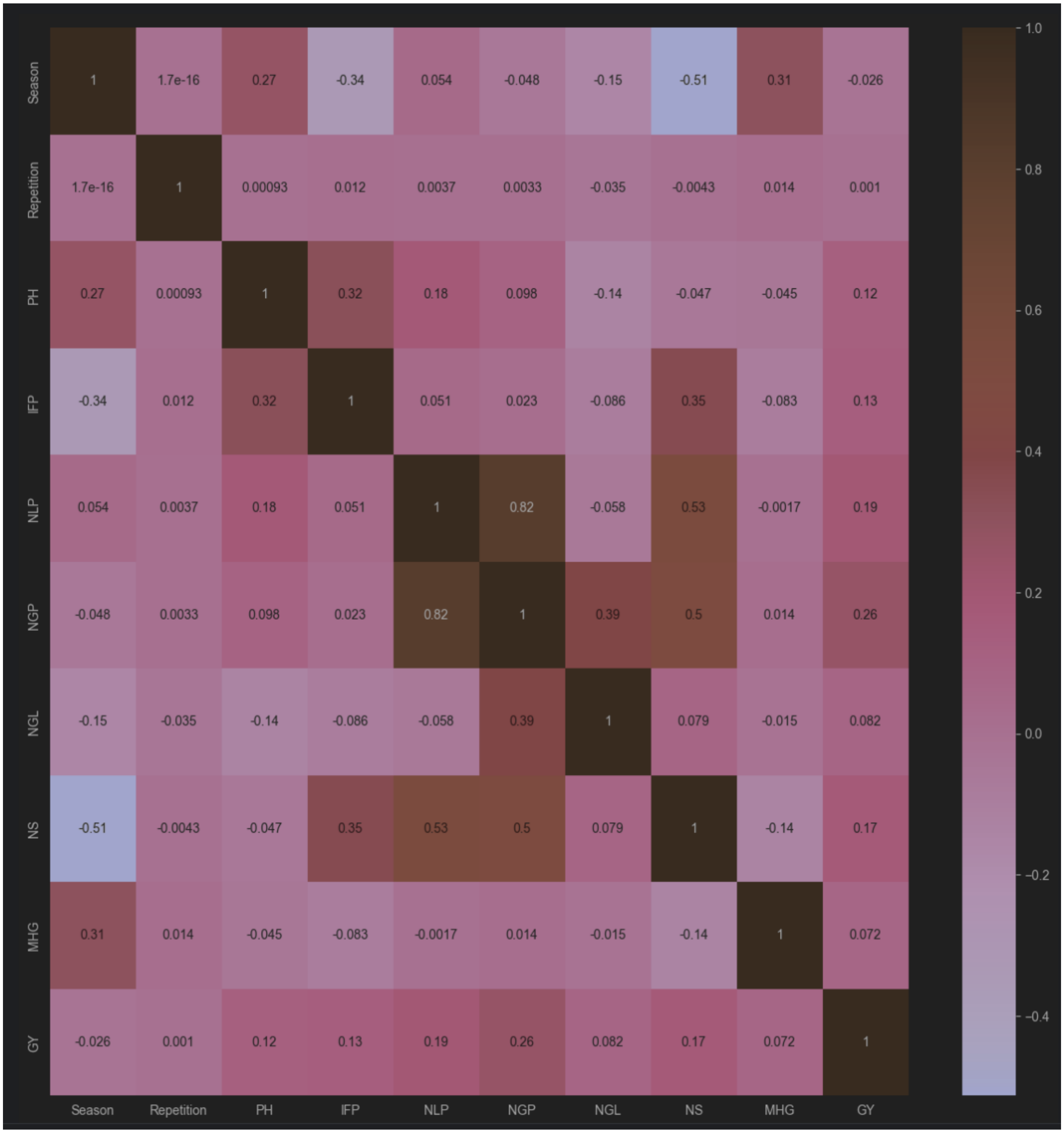
2.5. Detecting and removing outliers. In this step, outliers are identified and removed from the dataset to enhance the reliability of the regression model. The process involves visualizing the data using a boxplot to identify potential outliers and subsequently applying a filtering mechanism to exclude them



2.6. Correlation analysis. In this section, a correlation analysis is presented. The analysis, visualized through a heatmap, aims to provide insights into the correlation of Thousand Seed Weight (MHG) and Grain Yield (GY) with other variables.

2.6.1. Thousand Seed Weight (MHG) exhibits a moderate positive correlation coefficient of 0.31 with the variable 'Season', suggesting a potential seasonal influence on seed weight. Across all other variables, MHG demonstrates negligible correlations, implying minimal associations between MHG and the remaining variables in the dataset.

2.6.2. Grain Yield (GY) demonstrates a low to moderate positive correlation of 0.26 with the number of grains per plant (NGP), indicating a potential low relationship between grain yield and the number of grains per plant.

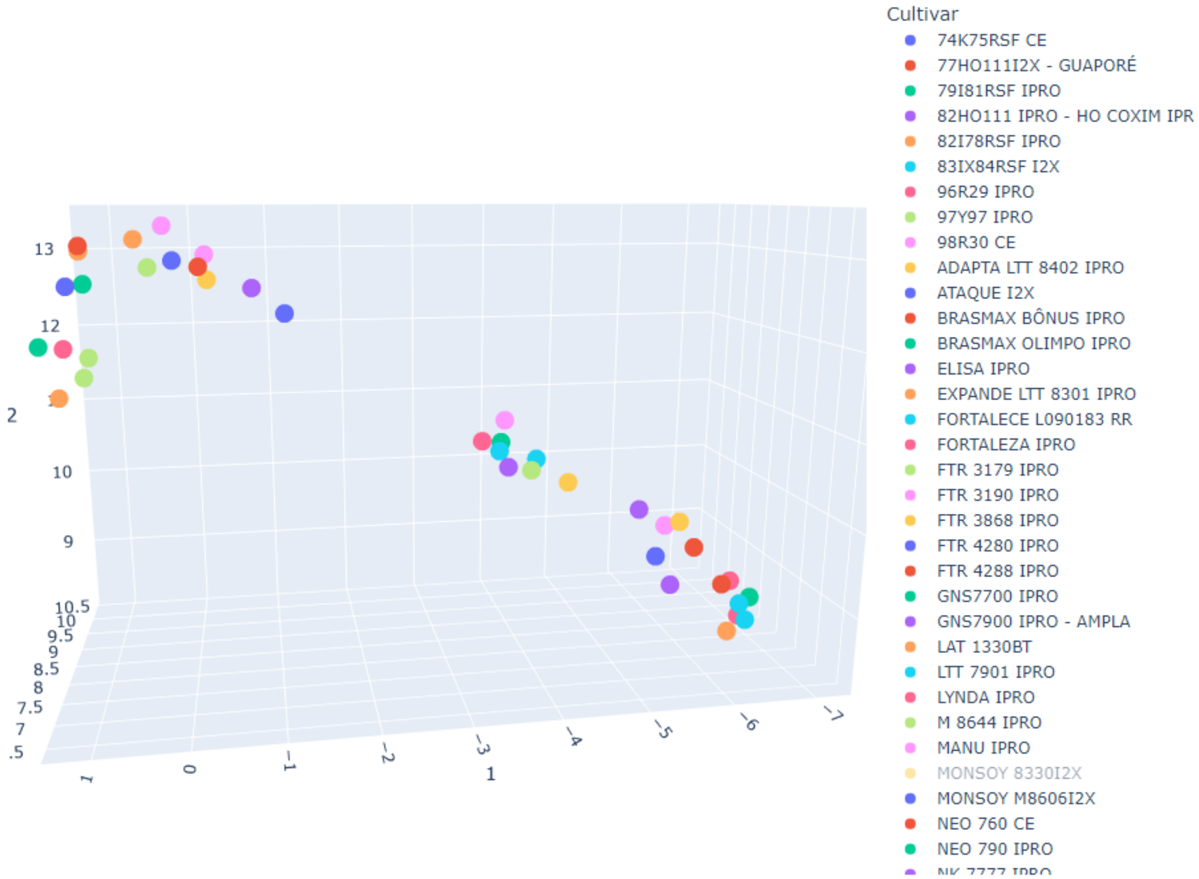


	Season	Repetition	PH	IFP	NLP	NGP	NGL	NS	MHG	GY
Season	1	1.7e-16	0.27	-0.34	0.054	-0.048	-0.15	-0.51	0.31	-0.026
Repetition	1.7e-16	1	0.00093	0.012	0.0037	0.0033	-0.035	-0.0043	0.014	0.001
PH	0.27	0.00093	1	0.32	0.18	0.098	-0.14	-0.047	-0.045	0.12
IFP	-0.34	0.012	0.32	1	0.051	0.023	-0.086	0.35	-0.083	0.13
NLP	0.054	0.0037	0.18	0.051	1	0.82	-0.058	0.53	-0.0017	0.19
NGP	-0.048	0.0033	0.098	0.023	0.82	1	0.39	0.5	0.014	0.26
NGL	-0.15	-0.035	-0.14	-0.086	-0.058	0.39	1	0.079	-0.015	0.082
NS	-0.51	-0.0043	-0.047	0.35	0.53	0.5	0.079	1	-0.14	0.17
MHG	0.31	0.014	-0.045	-0.083	-0.0017	0.014	-0.015	-0.14	1	0.072
GY	-0.026	0.001	0.12	0.13	0.19	0.26	0.082	0.17	0.072	1

3. Clustering.

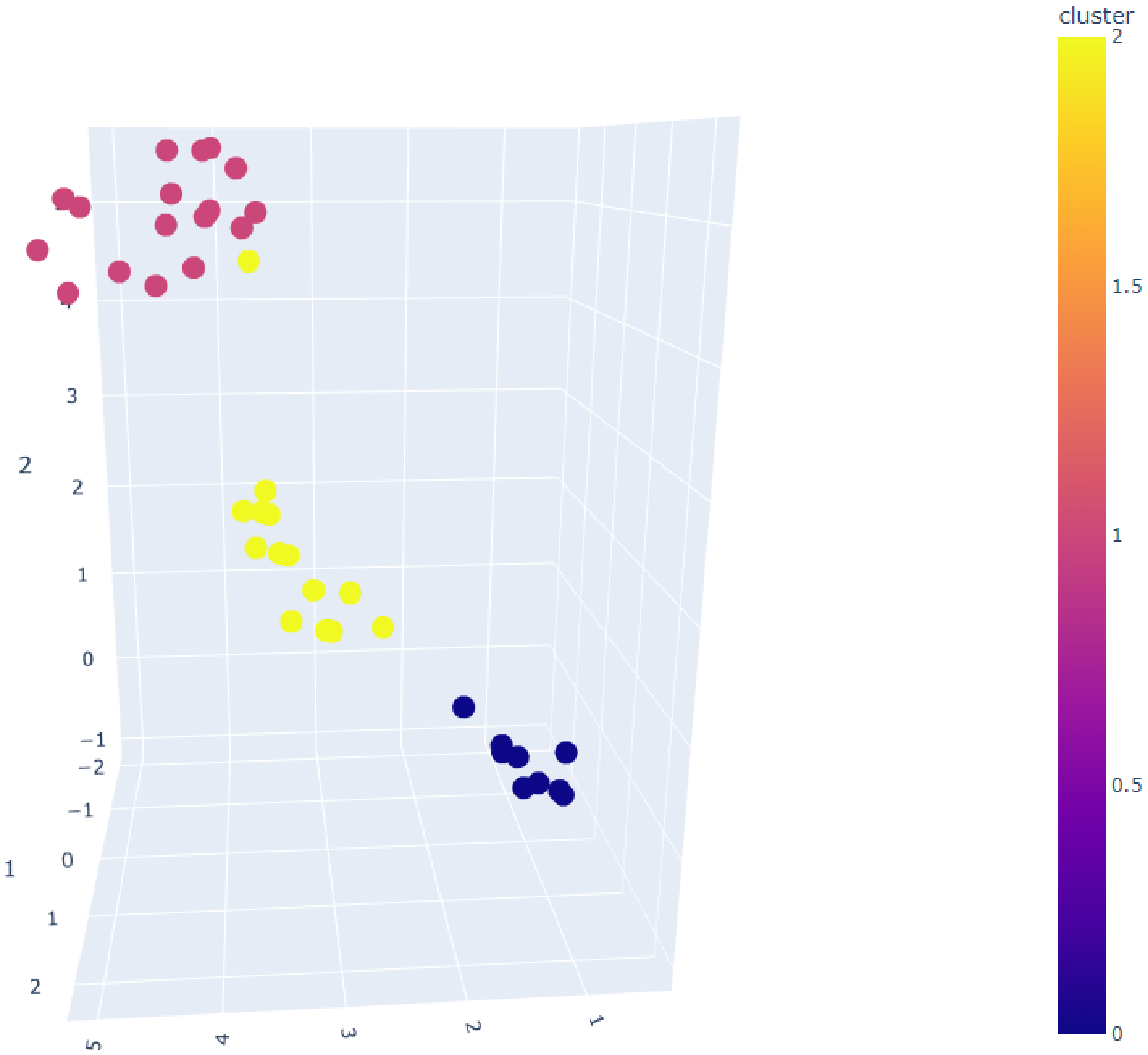
3.1. Preprocessing. In this step I transformed the initial dataset, which comprised 320 entries divided into two seasons with four repetitions each, into a condensed dataset containing 40 entries, with one entry per cultivar. The transformation process entails computing average values for each cultivar across both seasons and all repetitions. By consolidating the data in this manner, the resultant dataset provides a representative snapshot of each cultivar's traits.

3.2. Visualize the preprocessed data in 3D using dimensionality reduction. In this step I used dimensionality reduction techniques to visualize the condensed cultivar dataset in 3D space.

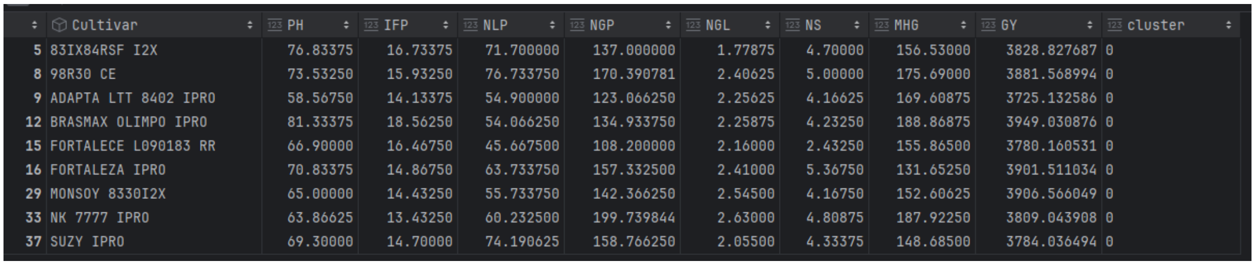


3.3. K-Means Clustering. In this step, the K-means clustering algorithm was applied to partition the condensed cultivar dataset into three distinct groups.

3.3.1. Visualization of clusters in 3D. The following image showcases the clusters visualized in 3D space, using dimensionality reduction techniques.



3.3.2. Cluster data. In the following image, data from a cluster is presented. Notably, we can observe close values for Grain Yield (GY) among cultivars within a cluster.



	Cultivar	PH	IFP	NLP	NGP	NGL	NS	MHG	GY	cluster
5	83IX84RSF I2X	76.83375	16.73375	71.700000	137.000000	1.77875	4.70000	156.53000	3828.827687	0
8	98R30 CE	73.53250	15.93250	76.733750	170.390781	2.40625	5.00000	175.69000	3881.568994	0
9	ADAPTA LTT 8402 IPRO	58.56750	14.13375	54.900000	123.066250	2.25625	4.16625	169.60875	3725.132586	0
12	BRASMAX OLIMPO IPRO	81.33375	18.56250	54.066250	134.933750	2.25875	4.23250	188.86875	3949.030876	0
15	FORTALECE L090183 RR	66.90000	16.46750	45.667500	108.200000	2.16000	2.43250	155.86500	3780.160531	0
16	FORTALEZA IPRO	70.83375	14.86750	63.733750	157.332500	2.41000	5.36750	131.65250	3901.511034	0
29	MONSOY 8330I2X	65.00000	14.43250	55.733750	142.366250	2.54500	4.16750	152.60625	3906.566049	0
33	NK 7777 IPRO	63.86625	13.43250	60.232500	199.739844	2.63000	4.80875	187.92250	3809.043908	0
37	SUZY IPRO	69.30000	14.70000	74.190625	158.766250	2.05500	4.33375	148.68500	3784.036494	0

4. Regression

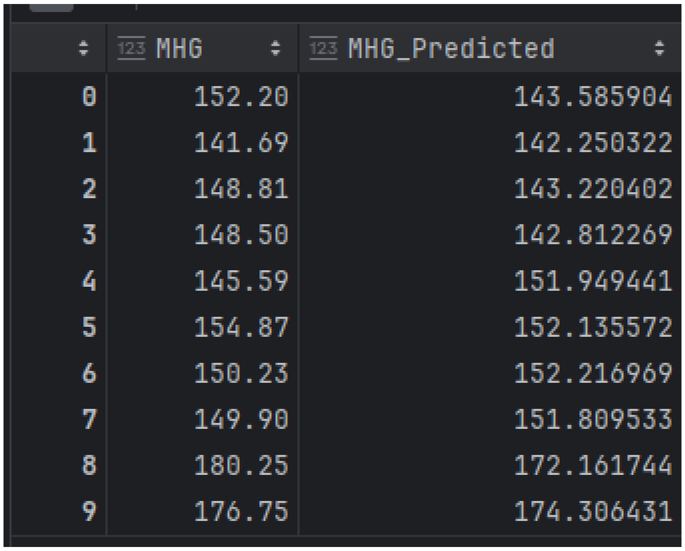
4.1. Introduction. In this section, the objective is to develop a regression model to predict the Thousand Seed Weight (MHG) from the initial dataset, trained on a synthetic dataset generated based on the initial dataset.

4.2. Generate synthetic data. In this step, we generate the synthetic dataset by using the mean and the standard deviation of each variable. We draw random samples from a normal distribution with means and standard deviations matching those estimated from the initial dataset.

4.3. Preprocess synthetic data. In this step, the synthetic dataset is preprocessed which involves one-hot encoding of categorical variables, specifically the ‘Cultivar’ and ‘Season’ columns.

4.4. Train the model. In this phase, a regression model is trained to predict the Thousand Seed Weight (MHG) based on the preprocessed synthetic dataset. A Linear Regression model was used. The performance metrics obtained from the regression model evaluation are as follows: Mean Squared Error (MSE) = ~58 and R-squared Score (R2) = ~0.86. The obtained MSE indicates the average squared difference between the predicted MHG values and the actual MHG values in the test set. A lowe MSE signifies a better predictive accuracy and model performance. The R-squared score indicates that the linear regression model explains approximately 86% of the variance in the MHG values, suggesting that the model provides a good fit to the data. Overall, the regression model does well in predicting MHG.

4.5. Prediction of MHG on initial dataset. In this final step, the trained model is utilized to predict the Thousand Seed Weight (MHG) values on the initial dataset. The following performance metrics are obtained from the prediction: MeanSquaredError (MSE) = ~52 and R-squared Score (R2) = ~0.86. These metrics indicate that the model does well on capturing the variability in MHG values. The following image showcases the real MHG and the predicted MHG:



	MHG	MHG_Predicted
0	152.20	143.585904
1	141.69	142.250322
2	148.81	143.220402
3	148.50	142.812269
4	145.59	151.949441
5	154.87	152.135572
6	150.23	152.216969
7	149.90	151.809533
8	180.25	172.161744
9	176.75	174.306431