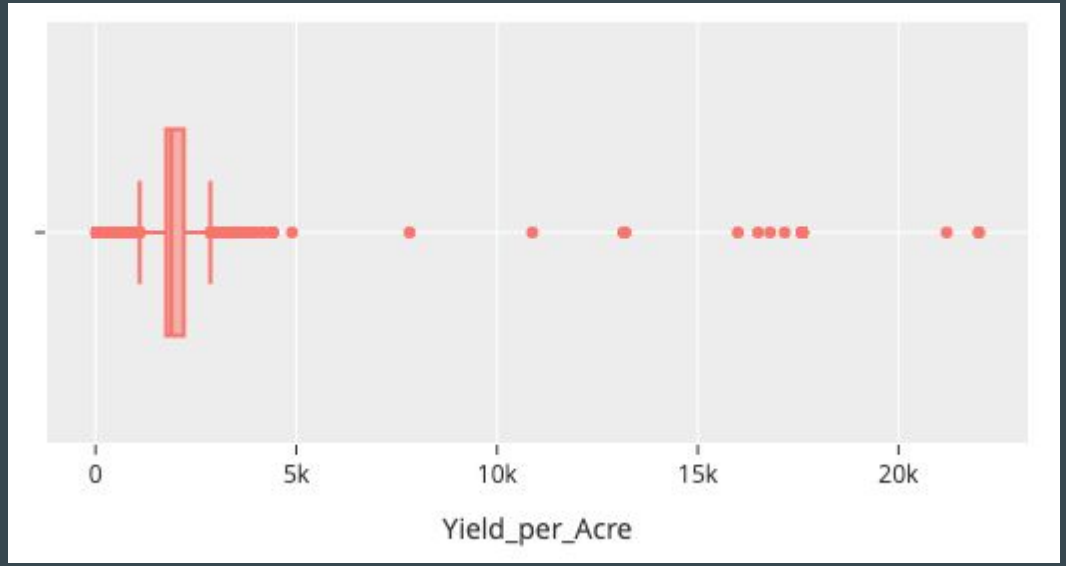
EDA, correlation analysis & clustering

Week 1

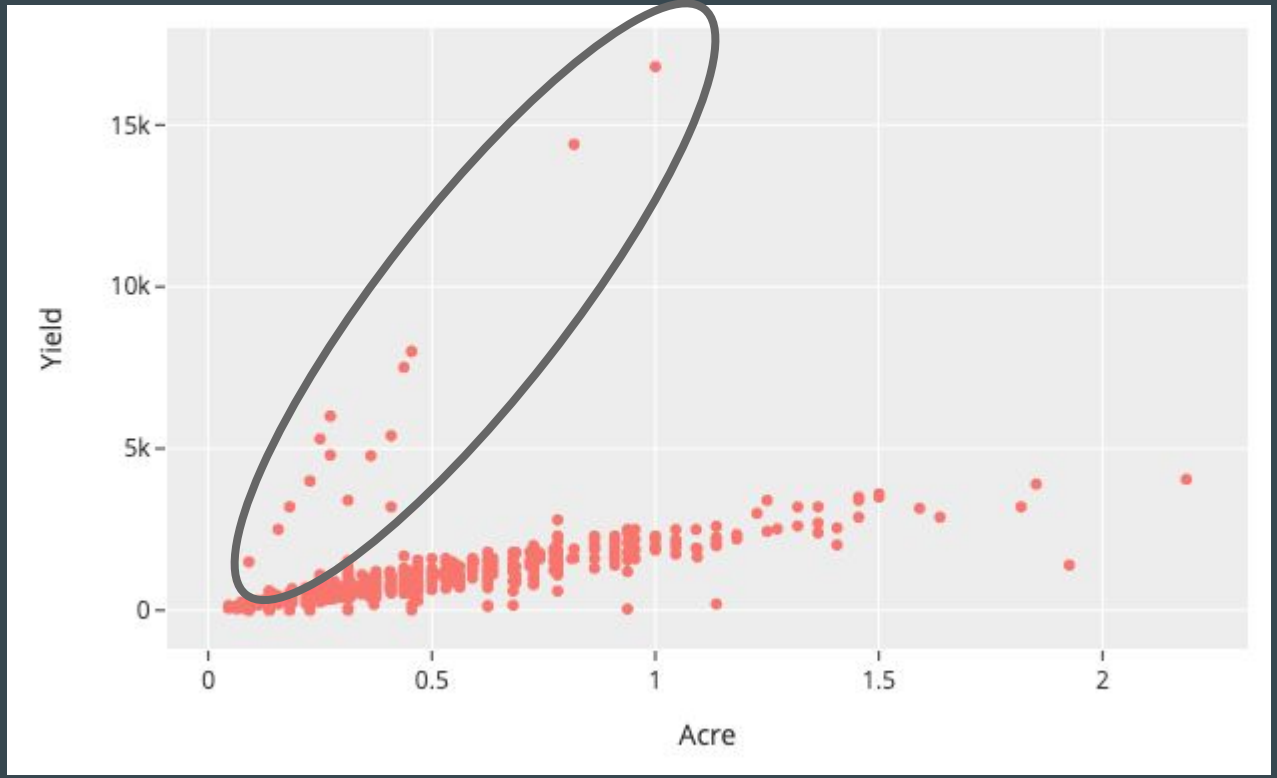
Yield-per-acre

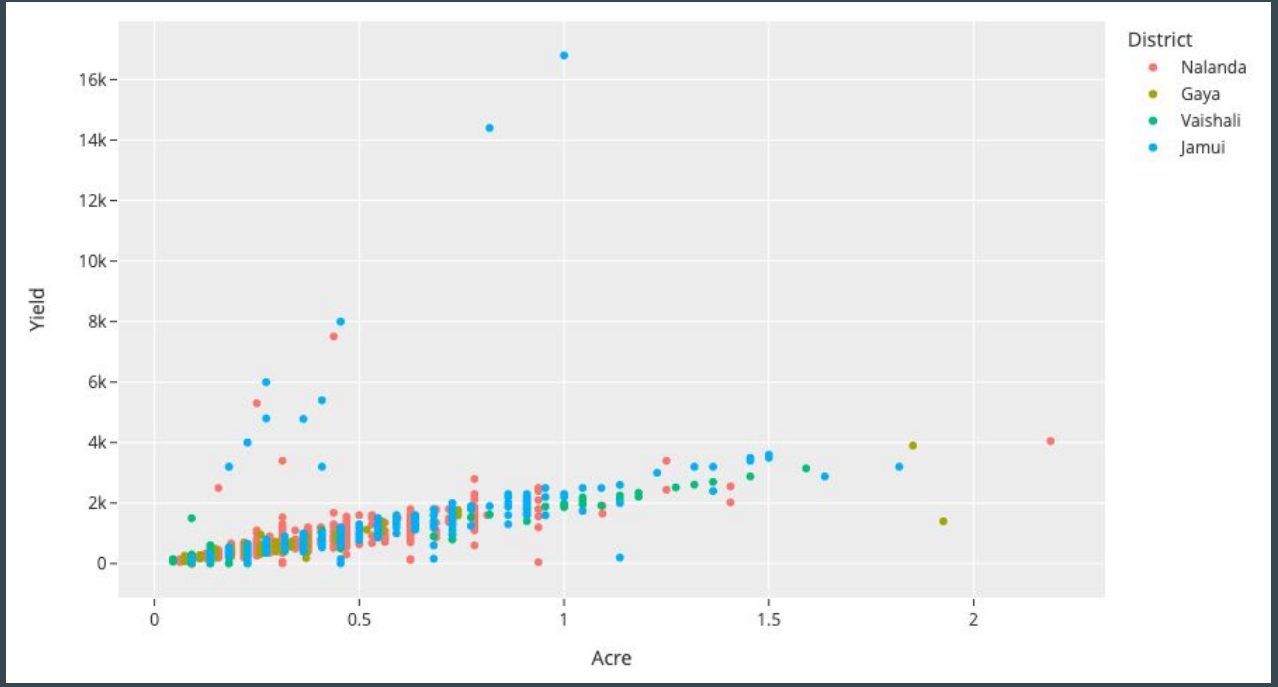


did the same for all variables marked by Alice

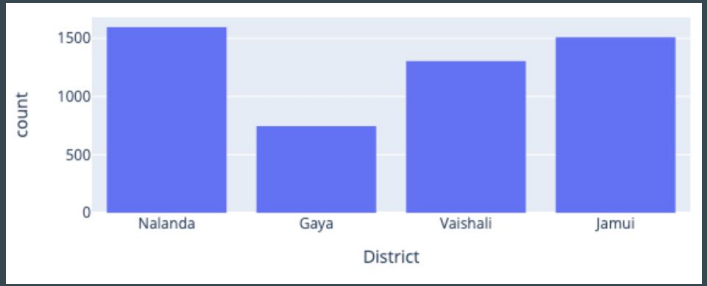
Yield & Acre scatter plot

0.68 correlation

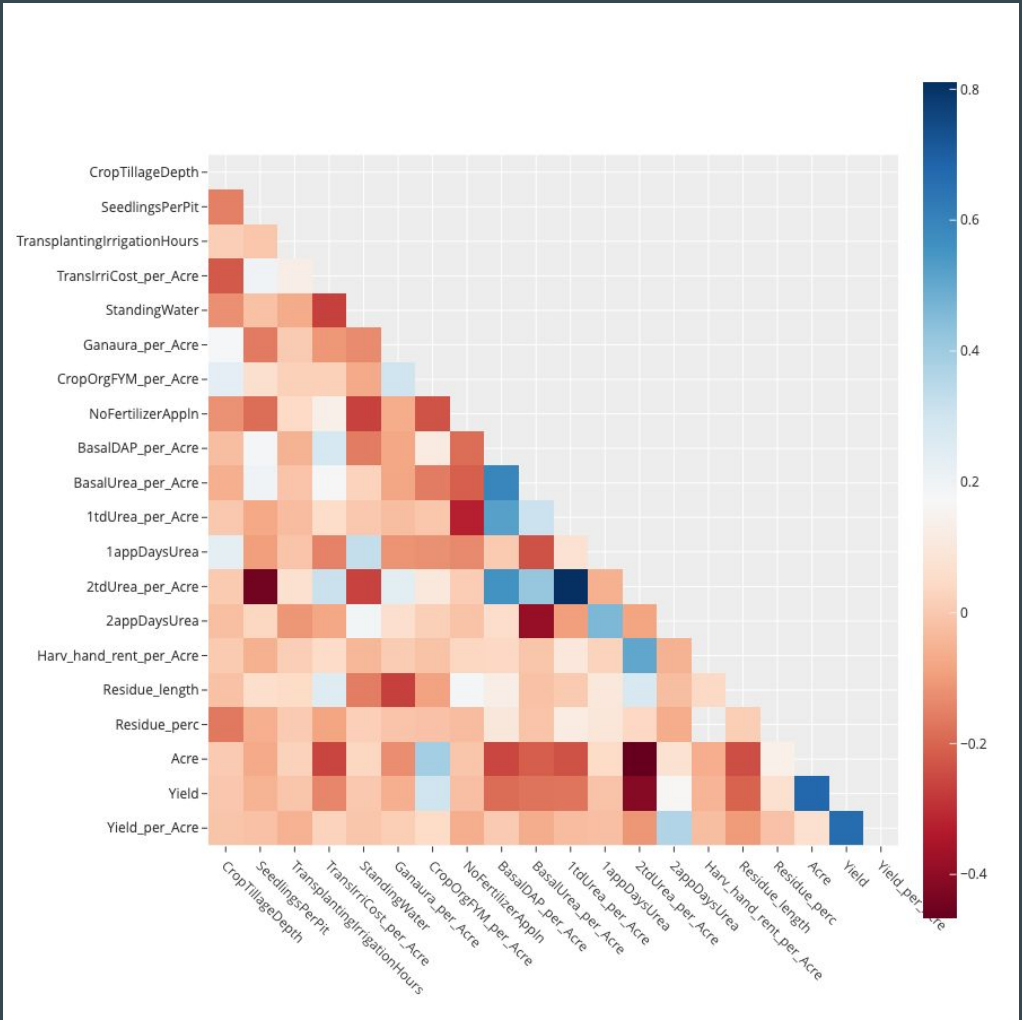




Overall district representation →



1. Correlation Analysis


CropTillageDepth
SeedlingsPerPit
TransplantingIrrigationHours
TransIrriCost_per_Acre
StandingWater
Ganaura_per_Acre
CropOrgFYM_per_Acre
NoFertilizerAppln
BasalDAP_per_Acre
BasalUrea_per_Acre
1tdUrea_per_Acre
1appDaysUrea
2tdUrea_per_Acre
2appDaysUrea
Harv_hand_rent_per_Acre
Residue_length
Residue_perc
Acre
Yield
Yield_per_Acre
0.8
0.6
0.4
0.2
0
−0.2
−0.4

Some interesting correlations...

- Small negative correlation beween 2tdUrea_per_Acre and Yield_per_Acre (r = -0.11)
- Small positive correlation between 2tdUrea_per_Acre and Residue_length (r = 0.27)
    - Residue_length is negatively correlated with Yield_per_Acre (r = -0.10)
- Moderate positive correlation between 2tdUrea_per_Acre and Harv_hand_rent_per_Acre (r = 0.51)
- Small positive correlation between 2appDaysUrea and Yield_per_Acre (0.37)
- Moderate negative correlation beween 2tdUrea_per_Acre and SeedlingsPerPit (r = -0.45) → could indicate different types of crops?
- There is only a 0.07 correlation between Acre and Yield_per_Acre

(full correlation map in the Google Sheets & on Git)

2. Comparing groups on outcomes

t-tests, ANOVAs & co

T-test results

Effect of x variable on Yield:

- **Harv_method** (hand vs. machine)
  - on Yield: t = -4.27, p = 0.00002, cohen's d = 0.29 → Small effect size of harvesting method on yield
  - on Yield_per_Acre: t= 0.87, p = 0.38, cohen's d = 0.05 → No effect of harvesting method on yield/acre
- **Threshing_method** (hand vs. machine)
  - on Yield: t = -2.24, p = 0.025, cohen's d = 0.07 → Very small effect size of threshing method on yield
  - on Yield_per_Acre: t = -3.98, p = 0.00007, cohen's d = 0.13 → Very small effect size of... on yield/acre
- **Stubble_use** (plowed in soil vs. burned)
  - on Yield: t = -1.81, p = 0.07, cohen's d = 0.37 → not enough "burned" instances (only 24 rows) to get a significant p-value, but could potentially be a meaningful predictor?
  - on Yield_per_Acre: t = 1.78, p = 0.07, cohen's d = 0.37 → same

Note: also ran non-parametric Mann-Whitney U tests → same results

Districts

note: not all districts / blocks are equal in terms of average land size

District	mean	median	std	count
Gaya	0.27	0.22	0.17	571
Jamui	0.34	0.23	0.22	1126
Nalanda	0.33	0.31	0.18	1193
Vaishali	0.20	0.14	0.21	980

Block	mean	median	std	count
Chehrakala	0.18	0.18	0.09	239
Garoul	0.48	0.23	0.40	134
Gurua	0.29	0.30	0.16	358
Jamui	0.31	0.23	0.19	626
Khaira	0.38	0.27	0.25	500
Mahua	0.15	0.14	0.09	607
Noorsarai	0.35	0.31	0.19	343
Rajgir	0.33	0.31	0.18	850
Wazirganj	0.24	0.19	0.17	213

Districts on Yield_per_Acre

The samples are not normally distributed and do not have equal variance → used Kruskal-Wallis test instead of ANOVA (tests for the median instead of the mean)

- Main effect is significant ($p<0.0001$)
- The only pairwise posthoc (Dunn's test) that isn't significant is Vaishali vs. Jamui; for the others, there is a significant difference in their yield per acre median.

Yield:

District	mean	median	std	count
Gaya	571.16	480.0	344.00	571
Jamui	730.27	450.0	966.98	1126
Nalanda	677.20	600.0	475.84	1193
Vaishali	350.52	250.0	413.60	980

Yield_per_Acre:

District	mean	median	std	count
Gaya	2071.66	2160.0	314.01	571
Jamui	2056.61	1760.0	1855.90	1126
Nalanda	2053.43	1920.0	1007.73	1193
Vaishali	1700.26	1760.0	869.27	980

Blocks on Yield_per_Acre

not normally distributed / no equal variance → Kruskal-Wallis test

- Main effect is significant ($p<0.0001$)
- 29/36 pairwise tests are statistically significant
- Again, could be due to different blocks cultivating different crops; or could be some other difference

	Chehrakala	Garoul	Gurua	Jamui	Khaira	Mahua	Noorsarai	Rajgir	Wazirganj
Chehrakala									
Garoul	0.00014								
Gurua	0.00000	0.00000							
Jamui	0.23512	0.00755	0.00000						
Khaira	0.00000	0.00002	0.11570	0.00000					
Mahua	0.00000	0.32974	0.00000	0.00000	0.00001				
Noorsarai	0.00000	0.00755	0.00207	0.00000	0.32974	0.04531			
Rajgir	0.00000	0.00008	0.00823	0.00000	0.45791	0.00001	0.45791		
Wazirganj	0.00000	0.00000	0.00775	0.00000	0.00000	0.00000	0.00000	0.00000	

Yield_per_Acre:

Block	mean	median	std	count
Chehrakala	1632.47	1650.00	313.21	239
Garoul	1807.06	1870.00	303.48	134
Gurua	2042.02	2106.00	335.56	358
Jamui	2098.66	1760.00	2469.63	626
Khaira	2003.96	1980.00	348.11	500
Mahua	1703.37	1833.33	1075.93	607
Noorsarai	1989.87	1920.00	634.26	343
Rajgir	2079.08	1920.00	1123.14	850
Wazirganj	2121.46	2160.00	267.42	213

Method of transplantation (CropEstMethod) on Yield_per_Acre

Kruskal-Wallis test

- Main effect is significant ($p<0.0001$)
- All methods have significantly different yield_per_acre medians from one another ($p<0.001$)

Yield_per_Acre:

CropEstMethod	mean	median	std	count
Broadcasting	2364.62	2560.0	314.49	83
LineSowingAfterTillage	1651.89	1664.0	411.73	206
Manual_PuddledLine	2042.47	1980.0	1014.73	235
Manual_PuddledRandom	1971.94	1890.0	1300.39	3346

TransplantingIrrigationSource on Yield_per_Acre

Kruskal-Wallis test

- Significant difference between Rainfed and TubeWell (p=0.02)

(but basically doesn’t really matter for yields, which makes sense)

TransplantingIrrigationPowerSource on Yield_per_Acre

- Also doesn’t matter for yields

PCropSolidOrgFertAppMethod on Yield_per_Acre

- Main effect is significant ($p<0.00001$)
- Significant difference in yields_per_acre median between Broadcasting and SoilApplied ($p<0.00001$)
- (the other 2 methods don't have enough data points)

Yield_per_Acre:

PCropSolidOrgFertAppMethod	mean	median	std	count
Broadcasting	1898.62	1760.00	2255.74	841
RootApplication	1932.22	2200.00	557.59	9
SoilApplied	2055.57	2055.24	581.99	1680
Spray	1186.33	1755.00	989.29	3

MineralFertAppMethod on Yield_per_Acre

- Main effect is significant ($p<0.00001$)
- No significant difference between RootApplication and Broadcasting ($p>0.05$), but significant between SoilApplied and Broadcasting ($p<0.0001$) and between SoilApplied and RootApplication ($p<0.05$)

Yield_per_Acre:

MineralFertAppMethod	mean	median	std	count
Broadcasting	1918.72	1833.33	1305.61	3214
RootApplication	1854.85	1876.67	444.30	18
SoilApplied	2217.07	2200.00	838.06	638

MineralFertAppMethod.1 (2nd dose) on Yield_per_Acre

- Main effect is significant ($p<0.00001$)
- All pairwise comparisons are significant ($p<0.01$)

Yield_per_Acre:

MineralFertAppMethod.1	mean	median	std	count
Broadcasting	1970.31	1907.45	1315.58	3288
RootApplication	2103.95	1706.67	2382.09	37
SoilApplied	2159.83	2200.00	466.72	64

Note: among the train set,

74% use the same method for the 1st and 2nd dose (mostly those using Broadcasting)

14% don't use the same method for the 1st and 2nd dose

12% don't have a 2nd dose

3. Identifying different crops?

unsupervised clustering attempt

Spectral clustering — Setup

Feature selection:

- took all variables indicated by Shaw, except for NursDetFactor and TransDetFactor (because from looking at the categories, I don't think it's actually helpful), and left out date variables.
- used /acre variables where needed

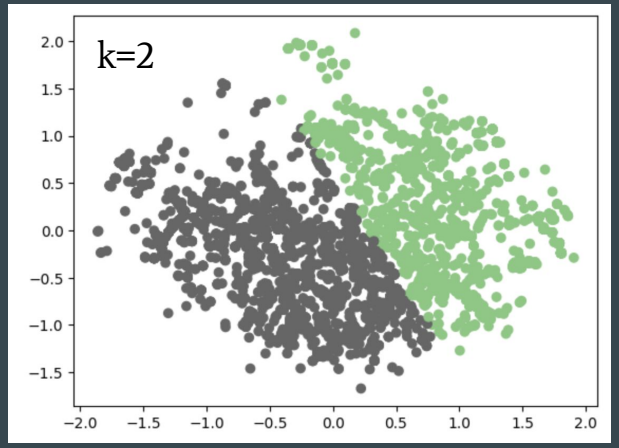
Cleaning:

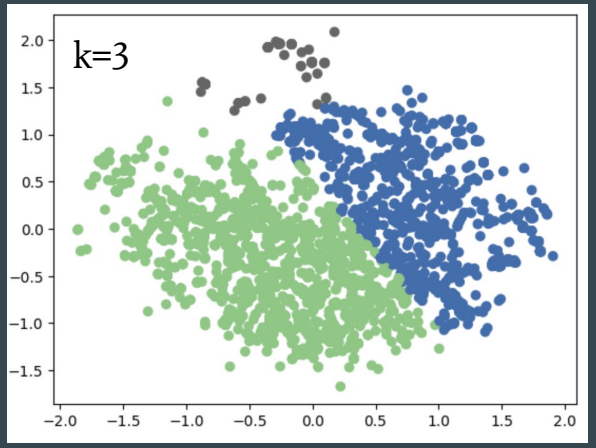
- outliers: capped some of the variables that had suspicious & extreme outliers (details in Gsheets doc)
- missing values:
  - for Ganaura_per_Acre & CropOrgFYM_per_Acre → replaced NaN with 0 (I'm assuming N/A means they didn't use any)
  - for SeedlingsPerPit, filled with the median (=2)
  - same for TransplantingIrrigationHours (median=4)

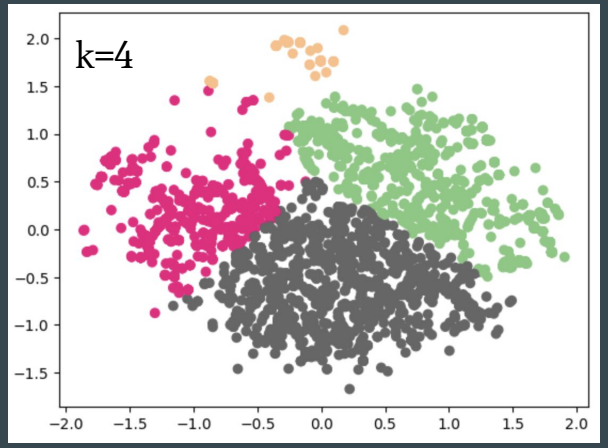
Pre-processing:

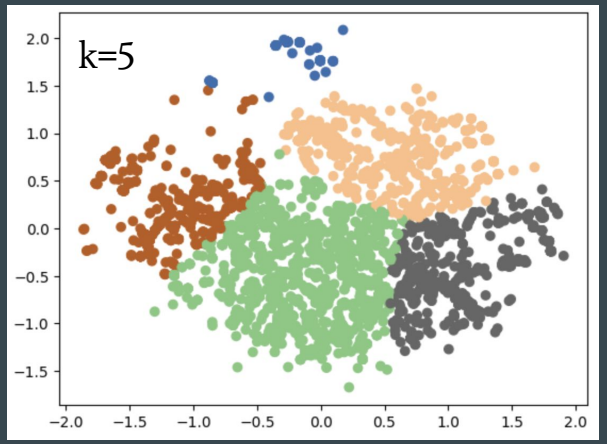
- 2 levels categorical variables → binary 0/1
- 3+ levels → dummies (after parsing the messy categorical variables)
- numerical → standard scaling & normalization

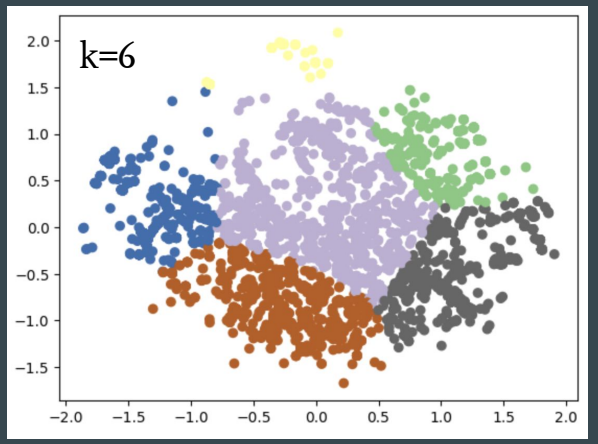
Spectral clustering — Results (with PCA to visualize)











Spectral clustering — Results

- Ran spectral clustering for k=2-5
- k=2
  - Cluster A: 4910
  - Cluster B: **250**
- k=3
  - Cluster A: 3907
  - Cluster B: **1004**
  - Cluster C: **249**
- k=4
  - Cluster A: 3837
  - Cluster B: **987**
  - Cluster C: **249**
  - Cluster D: **87**
- k=5
  - Cluster A: 2383
  - Cluster B: **2085**
  - Cluster C: **356**
  - Cluster D: **249**
  - Cluster E: **87**

- **Orange cluster**: match between all model (the same rows are systematically clustered together) (there's just 1 that jumps to the other group at k=3)
- **Cyan cluster**: split from k=2's cluster A, remains stable at k=4, and splits in 2 at k=5
- **White cluster**: split from k=2-3's cluster A

- k=2's Cluster A splits as k increases

Hypothesis?

- k=2 reveals Wheat vs. Rice, and k>2 differentiates between subtypes of either rice or wheat (maybe?)

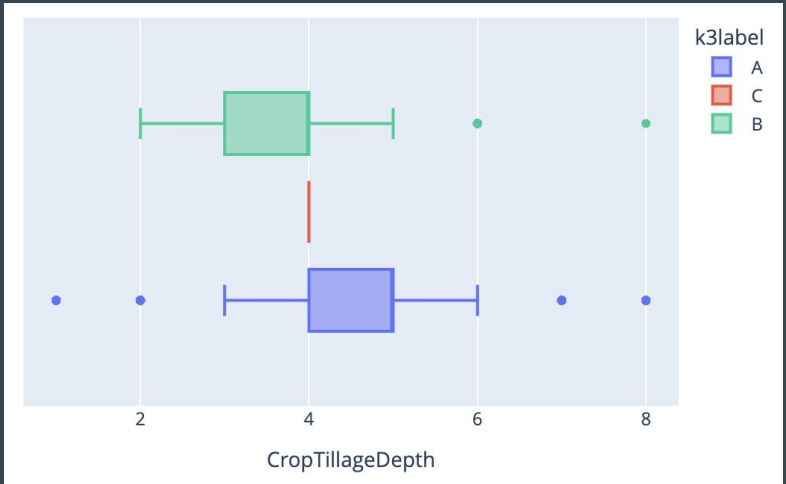
Spectral clustering — Visualizing clusters on different variables

(looked at the k=3 clusters)

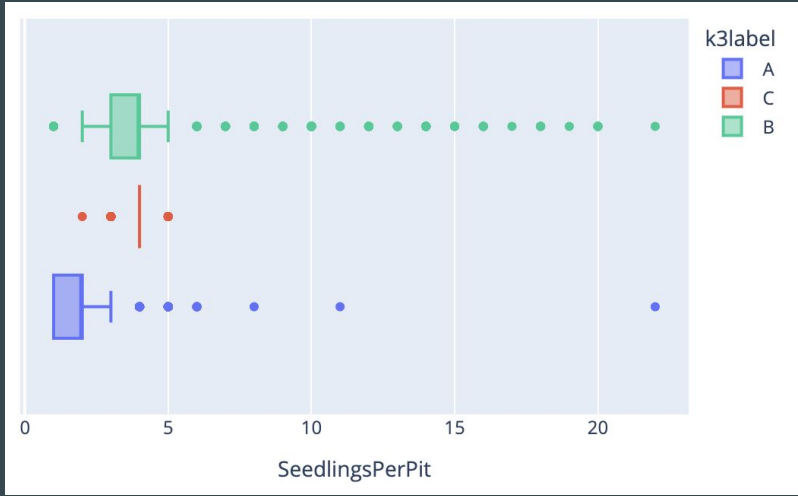
CropTillageDepth:

for wheat, the recommended tillage depth is 15cm (=6 inches, but idk if the variable is in inches); for rice, it's 17-20cm

→ all Cluster Cs have the same CropTillageDepth value (4), while Cluster A values range from 1 to 8 (median=4)



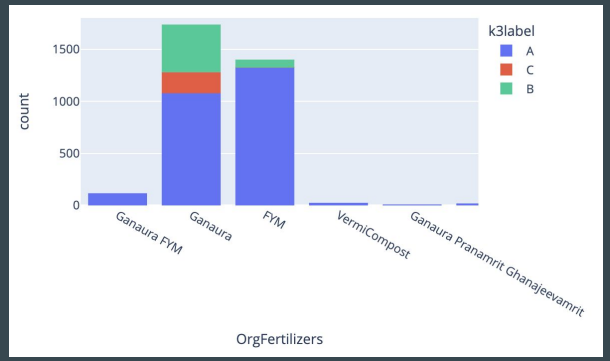
SeedlingsPerPit:

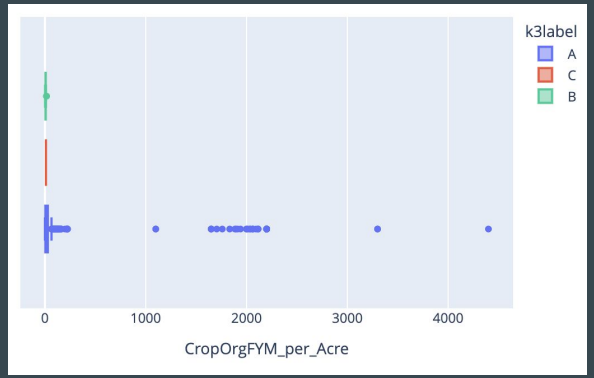


Spectral clustering — Visualizing clusters on different variables

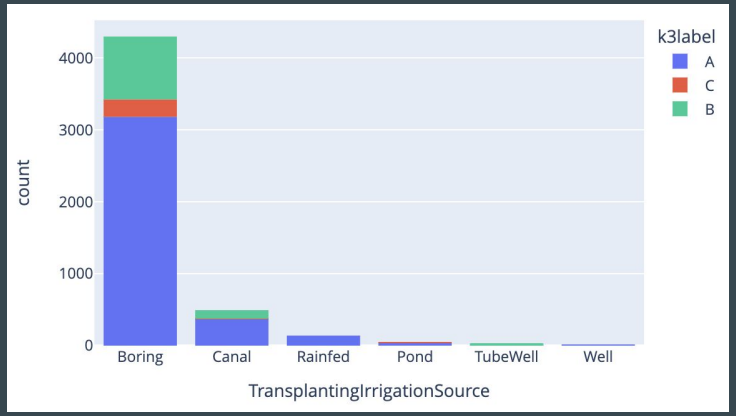
(looked at the k=3 clusters)

Fertilizers:





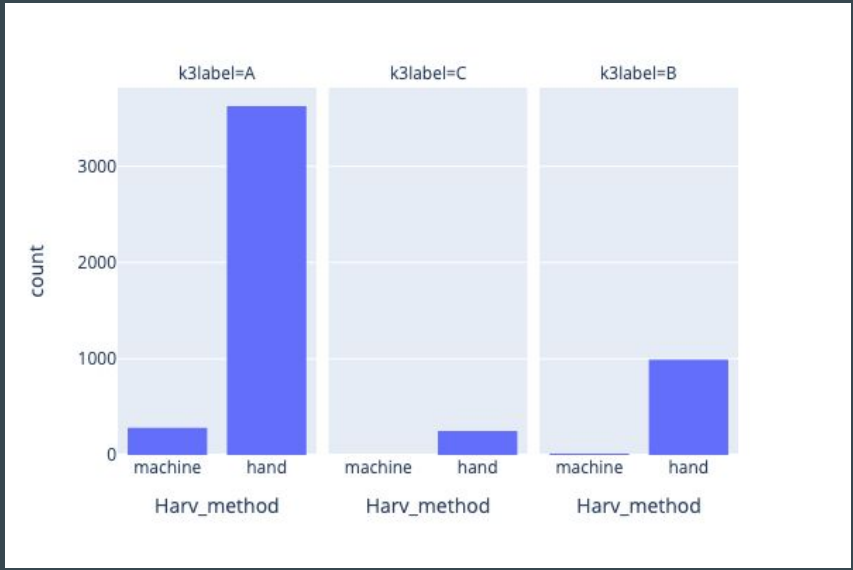
TransplantingIrrigationSource:



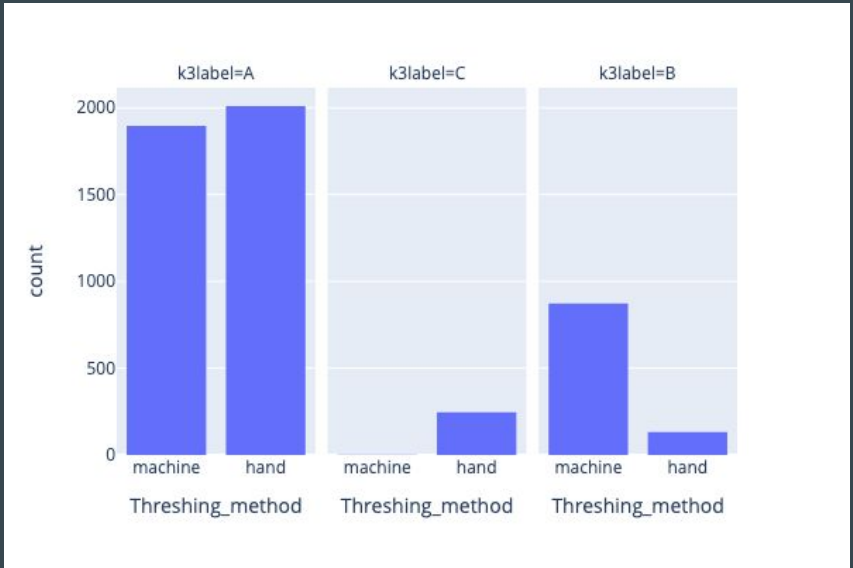
Spectral clustering — Visualizing clusters on different variables

(looked at the k=3 clusters)

Harvesting method



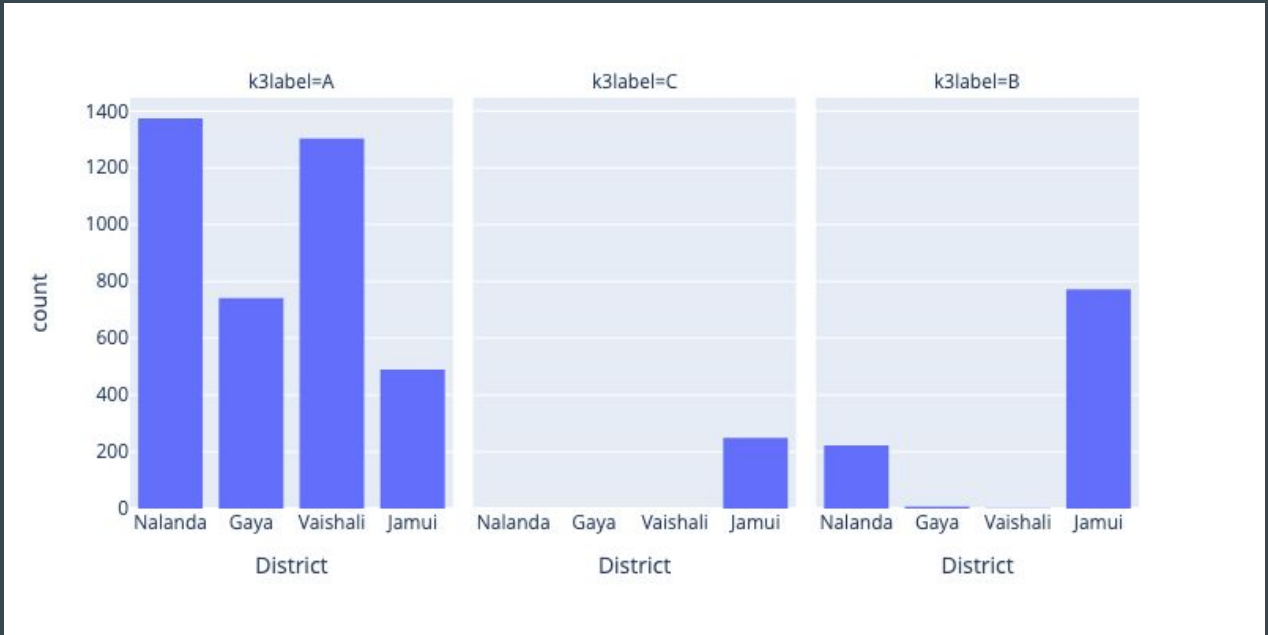
Threshing method



Spectral clustering — Visualizing clusters on different variables

(looked at the k=3 clusters)

Districts



Spectral clustering — Next steps?

- For evaluation / interpretation, which variables should we focus on? Which ones could be the clearest indicators of crop type?
    - comparing clusters on these variables to confirm that the clusters are capturing crop type and not something else?

- could compare results from a different model (e.g. DBSCAN)
- could refine feature selection to make sure we're capturing crop types and not adding distractors