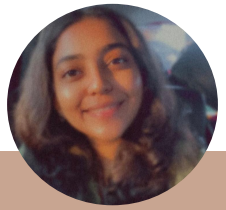
Beans to Bytes

Leveraging Technology for Agricultural Innovation



Project Presentation by Group 84: Harmeet Singh Rushaliben More Siraz Uz Zaman

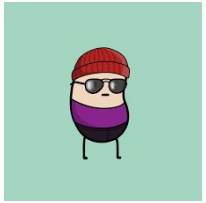


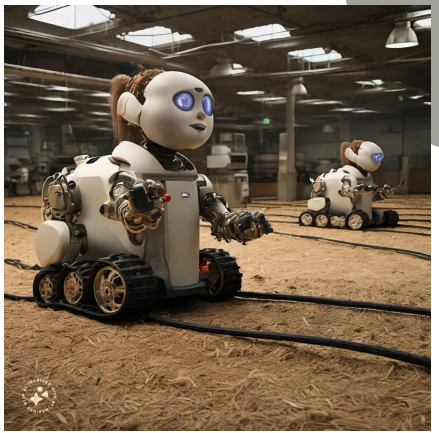
Table of contents

Why Beans?

BEANALYTICS
Don't Worry about Spilled Beans

According to the Food and Agriculture Organizations (FAO), beans are a staple food for more than 400 million people around the world. Not only are the consumption numbers staggering but the impact that bean cultivation has on global agriculture is also significant. It is noted that in the US alone, more than 1.5m acres of land is devoted to bean cultivation which leads to a production of 2.2b pounds of beans every year. Given the scale of the bean cultivation market and consumption, there is a potential to tap into this market for technologically advanced farming equipment manufacturers.

In today's times, large scale farming has adopted technology, starting from automated irrigation systems to drone harvesting techniques. Imagine a world where robots could identify a bean plant and nurture the bean's journey from correct nutrition at growing stage to automated sorting and packaging in large bean factories. This is what is at the heart of **Beanalytics LLC**, a silicon valley tech company founded to drive agricultural innovation beyond the unimaginable.



Source: Images generated by MetaAI

Objectives of Our Research

BEANALYTICS
Don't Worry about Spilled Beans

Bean Classification:

Leveraging advanced technologies, particularly through the application of unsupervised clustering. The primary objective is to categorize and comprehend inherent patterns within diverse bean varieties, employing machine learning algorithms to identify natural groupings of bean varieties.

Technologically Advanced Farming:

Beanalytics seeks to contribute to the development of technologically advanced farming equipment. This technology aims to discern, nurture, and process beans with unprecedented accuracy, validating its practicality for a future where automation seamlessly integrates with the intricate world of bean cultivation, ensuring optimal yields and sustainable agricultural practices.

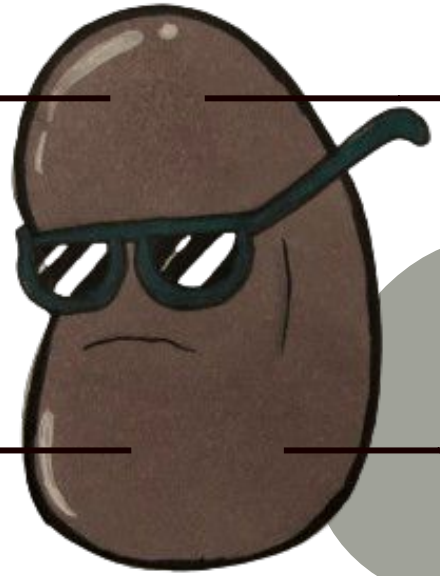
BEANALYTICS
Don't Worry about Spilled Beans

Data Exploration

BEANALYTICS
Don't Worry about Spilled Beans

13,911

Grain Records



16

Attributes

Such as Area, Eccentricity, Convex area, roundness, aspect ratio, among other

7

Classification Categories

This represents the different Bean Varieties (Seker, Barbunya, Bombay, Cali, Horoz, Sira, Dermason)

0

Missing Values

Insights from Data Exploration

BEANALYTICS
Don't Worry about Spilled Beans

Data Features

– 15 numerical features
– 1 Class feature (labels)

```
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Area             13611 non-null  int64
 1   Perimeter        13611 non-null  float64
 2   MajorAxisLength  13611 non-null  float64
 3   MinorAxisLength  13611 non-null  float64
 4   AspectRation     13611 non-null  float64
 5   Eccentricity     13611 non-null  float64
 6   ConvexArea       13611 non-null  int64
 7   EquivDiameter    13611 non-null  float64
 8   Extent           13611 non-null  float64
 9   Solidity         13611 non-null  float64
 10  roundness        13611 non-null  float64
 11  Compactness      13611 non-null  float64
 12  ShapeFactor1     13611 non-null  float64
 13  ShapeFactor2     13611 non-null  float64
 14  ShapeFactor3     13611 non-null  float64
 15  ShapeFactor4     13611 non-null  float64
 16  Class            13611 non-null  object
dtypes: float64(14), int64(2), object(1)
```

Data Labels

7 Distinct Bean Classifications

```
Bean Classes and Count of Occurences:


DERMASON    3546
SIRA        2636
SEKER       2027
HOROZ       1928
CALI        1630
BARBUNYA    1322
BOMBAY       522
Name: Class, dtype: int64
```

Insights from Data Exploration

BEANALYTICS
Don't Worry about Spilled Beans

Summary Statistics of Numeric Features

	Area	Perimeter	MajorAxisLength	MinorAxisLength	AspectRation	Eccentricity	ConvexArea	EquivDiameter	Extent	Solidity	roundness	Compactness	ShapeFactor1	ShapeFactor2	ShapeFa
count	13611.000000	13611.000000	13611.000000	13611.000000	13611.000000	13611.000000	13611.000000	13611.000000	13611.000000	13611.000000	13611.000000	13611.000000	13611.000000	13611.000000	13611.0
mean	53048.284549	855.283459	320.141867	202.270714	1.583242	0.750895	53768.200206	253.064220	0.749733	0.987143	0.873282	0.799864	0.006564	0.001716	0.6
std	29324.095717	214.289696	85.694186	44.970091	0.246678	0.092002	29774.915817	59.177120	0.049086	0.004660	0.059520	0.061713	0.001128	0.000596	0.0
min	20420.000000	524.736000	183.601165	122.512653	1.024868	0.218951	20684.000000	161.243764	0.555315	0.919246	0.489618	0.640577	0.002778	0.000564	0.4
25%	36328.000000	703.523500	253.303633	175.848170	1.432307	0.715928	36714.500000	215.068003	0.718634	0.985670	0.832096	0.762469	0.005900	0.001154	0.5
50%	44652.000000	794.941000	296.883367	192.431733	1.551124	0.764441	45178.000000	238.438026	0.759859	0.988283	0.883157	0.801277	0.006645	0.001694	0.6
75%	61332.000000	977.213000	376.495012	217.031741	1.707109	0.810466	62294.000000	279.446467	0.786851	0.990013	0.916869	0.834270	0.007271	0.002170	0.6
max	254616.000000	1985.370000	738.860154	460.198497	2.430306	0.911423	263261.000000	569.374358	0.866195	0.994677	0.990685	0.987303	0.010451	0.003665	0.9

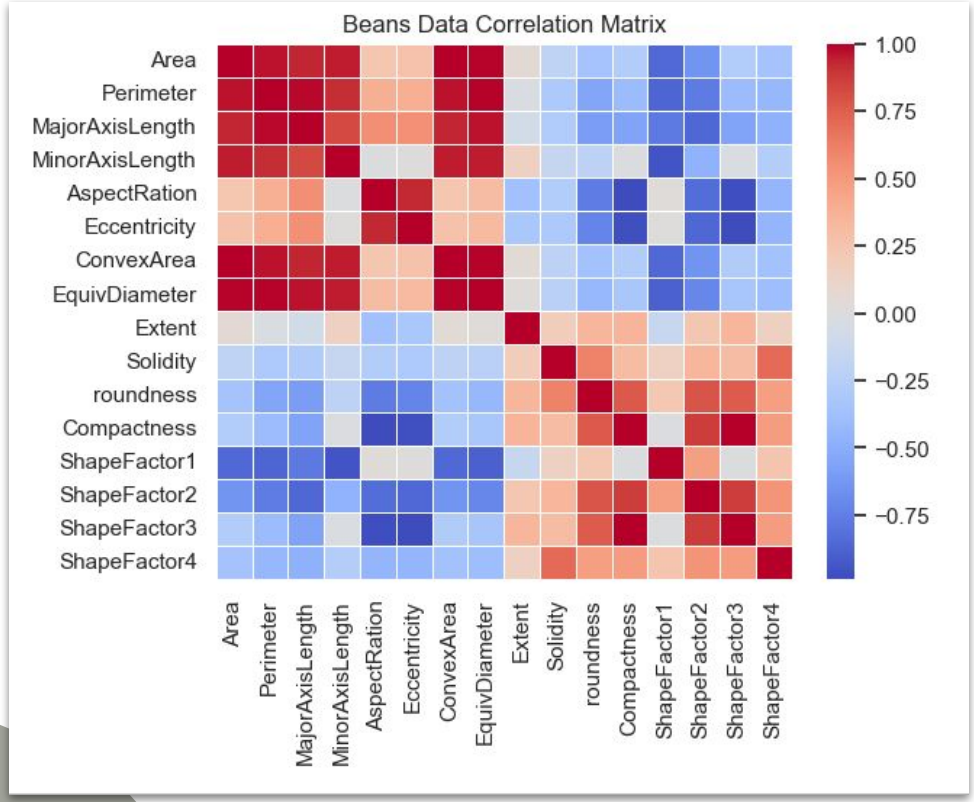
- The **Area** feature has a range of value from **20,420 to 254,616** pixels, suggesting significant variation in the size of the different beans. Similar interpretation can be made of the perimeter feature as it is also quite varied.
- The **average aspect ratio** of the beans is **1.58** which suggests that the beans are not very **elongated.** Although the **minimum and maximum values are 1.02 and 2.43 respectively,** which suggests that there are beans that are almost **circular** and those that have lengths twice the size of their width (and thus **more elongated**).
- The **average roundness of the beans** is approximately **0.87** which means that the beans in the dataset are **not perfectly circular** but close to a circular shape.
- Sufficient information to identify correct type of beans

BEANALYTICS
Don't Worry about Spilled Beans

Data Visualization

Spill the Beans!

Correlation Plot of Beans Data



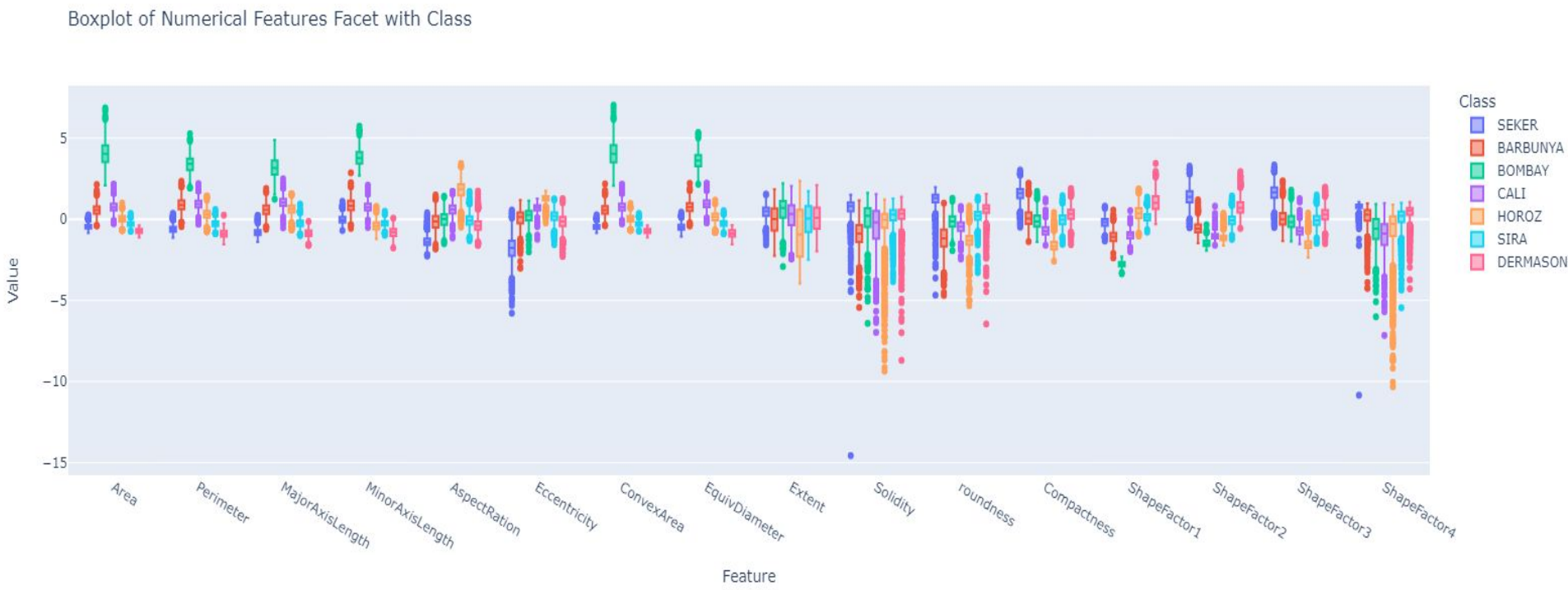
We observe:

- High correlation between bean dimensional features (eg Area, perimeter, axis lengths)
- There is a mix of both positively correlated and negatively correlated features.

We note that:

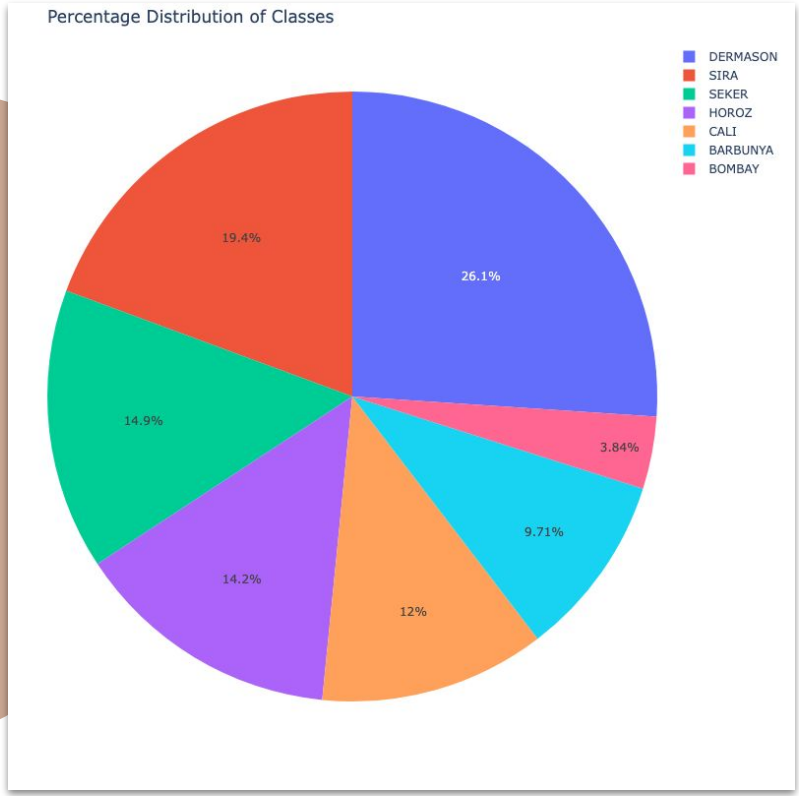
- All features are relevant and carry important information that could enable classification of beans.

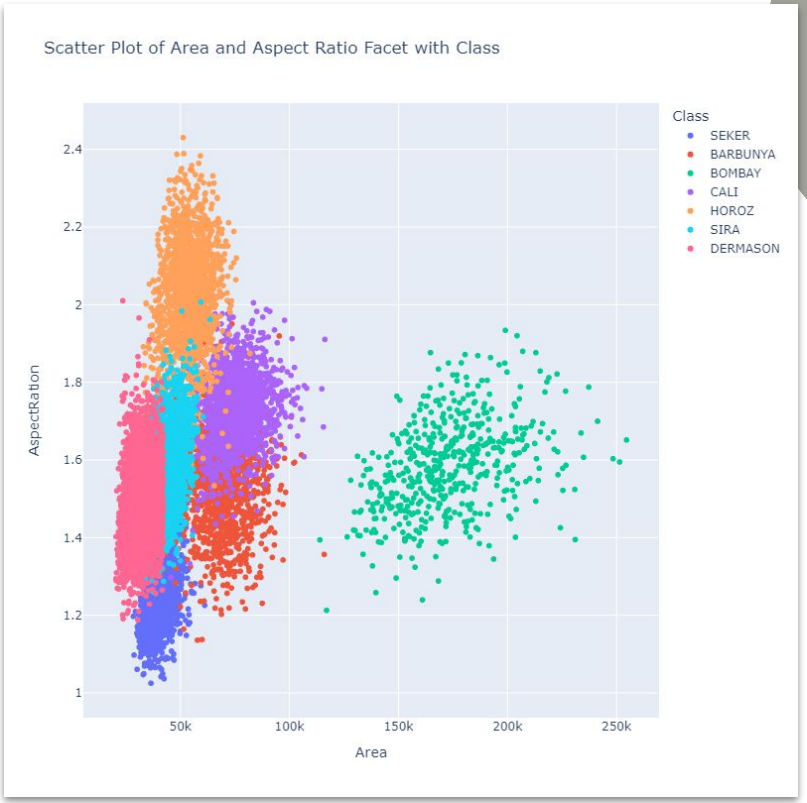
Boxplots for Beans Data by Class



DATA VISUALIZATION

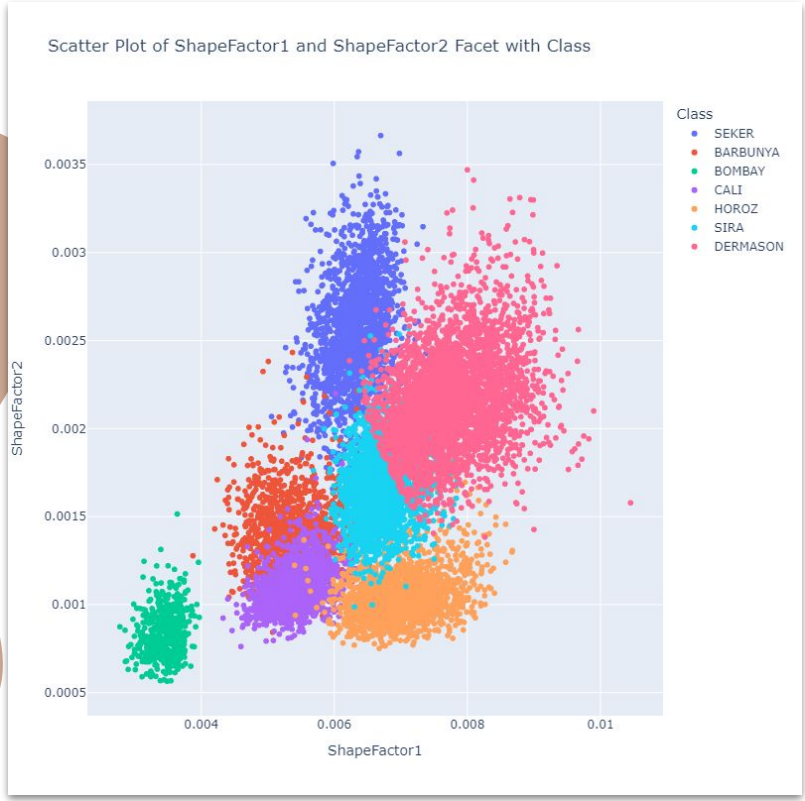
BEANALYTICS
Don't Worry about Spilled Beans

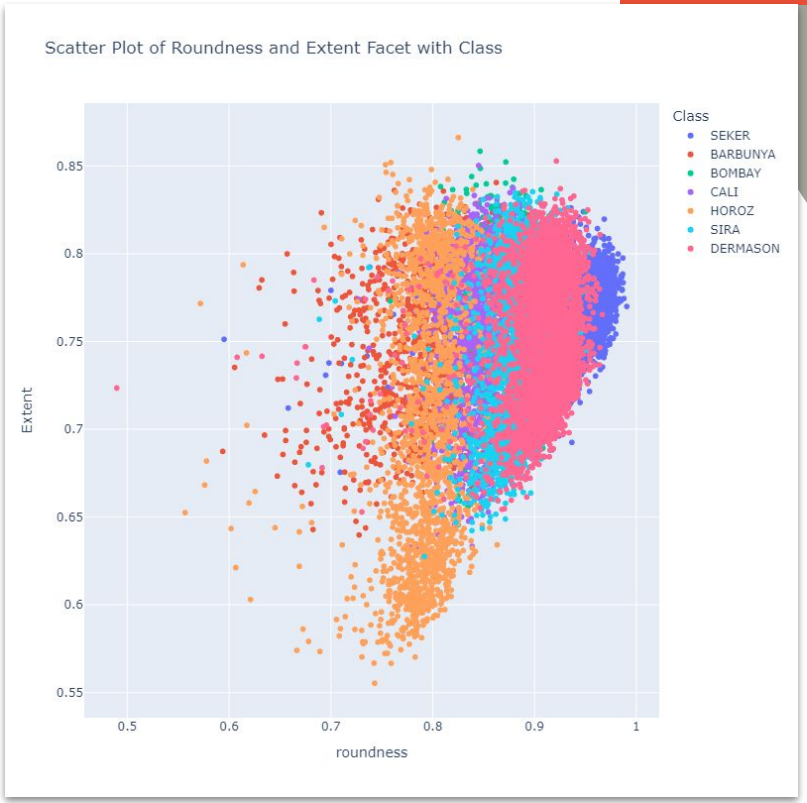


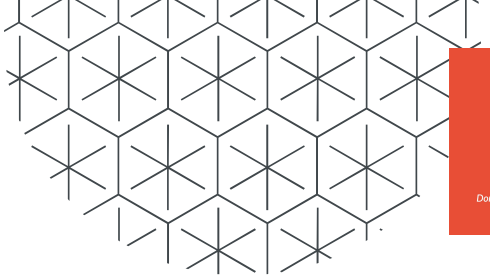


DATA VISUALIZATION

BEANALYTICS
Don't Worry about Spilled Beans







Feature Selection

Dimensionality Reduction Using PCA

- Data we have is important and needs to be retained
- Data that we have is highly correlated

Why PCA then?

- Retain all Features
- Features are highly correlated, PCA is effective in reducing multicollinearity
- PCA captures significant variability and reduces noise
- Model performance Vs Interpretability.. We care about the former

Dimensionality Reduction Using PCA

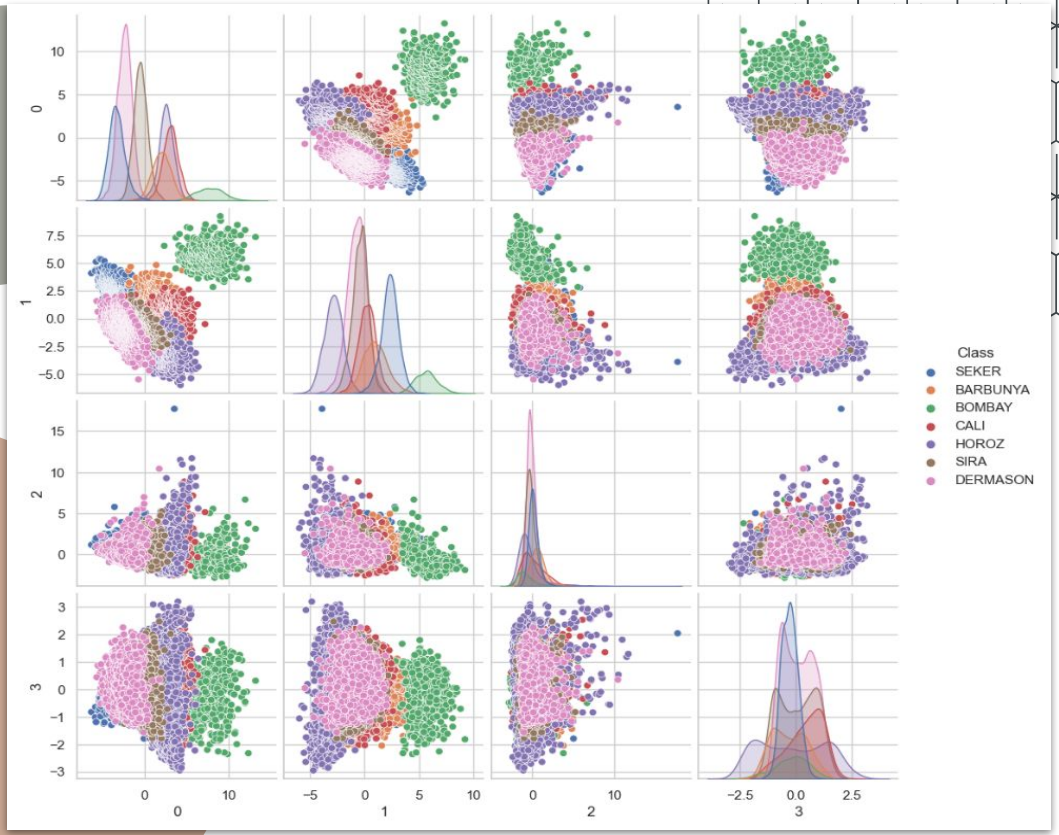
BEANALYTICS
Don't Worry about Spilled Beans

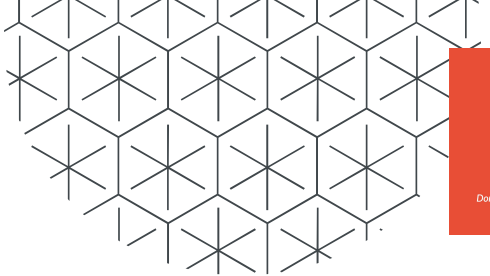
Criteria for selecting number of PCAS? – at least 95% of explained variance ratio

Result:

```
Principal Component 1: Explained Variance = 0.5547, Cumulative Variance = 0.5547
Principal Component 2: Explained Variance = 0.2643, Cumulative Variance = 0.8190
Principal Component 3: Explained Variance = 0.0801, Cumulative Variance = 0.8990
Principal Component 4: Explained Variance = 0.0511, Cumulative Variance = 0.9502
```

Pair Plot of PCAs





Unsupervised Clustering

K- Means Clustering

K - Means Clustering: Model Parameters

BEANALYTICS
Don't Worry about Spilled Beans

Parameters

Number of Clusters = **1 to 15 (loop)**

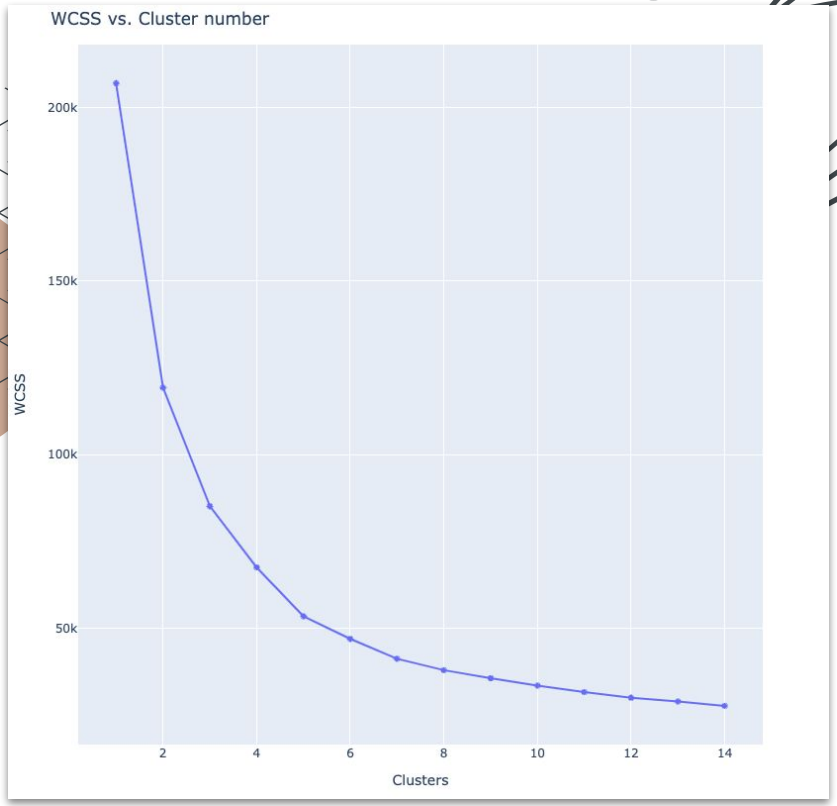
Number of iterations = **500**

Number of times the k-means algorithm will be run with different centroid seeds = **10**

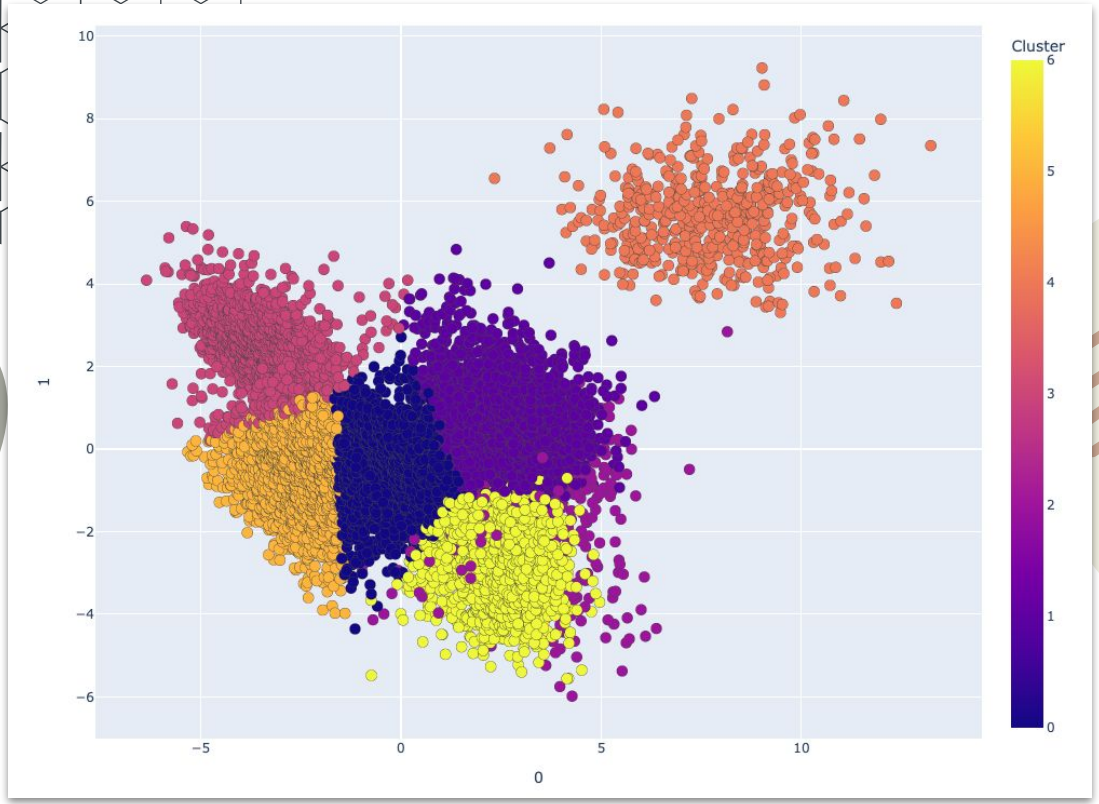
Evaluation Criteria: **WCSS to determine optimum number of clusters**

K - Means Clustering: WCSS

BEANALYTICS
Don't Worry about Spilled Beans



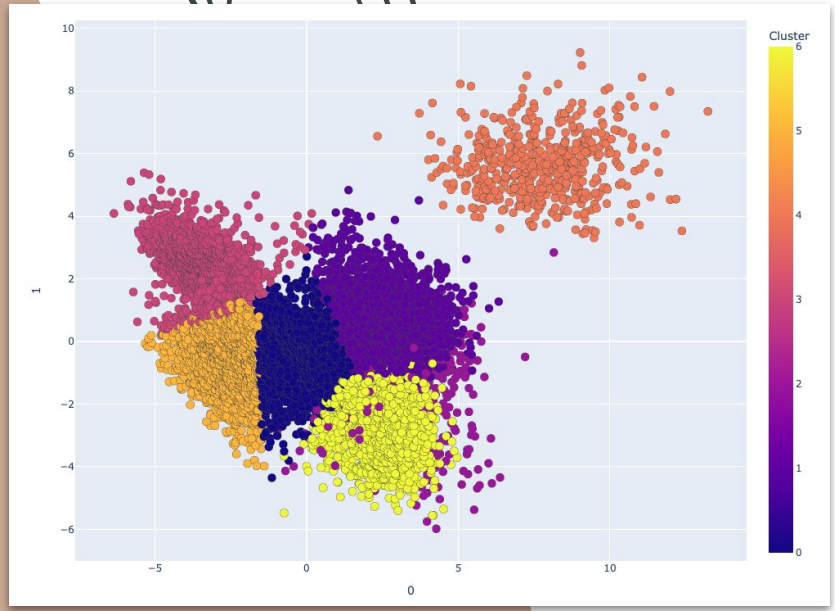
K - Means Clustering: Result



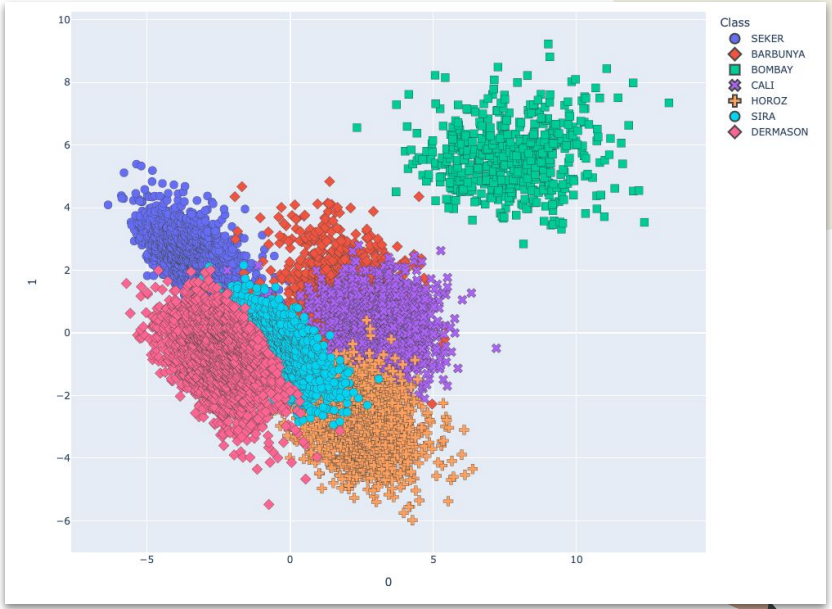
BEANALYTICS
Don't Worry about Spilled Beans

K- Means Clustering: Comparison with Labelled Data

BEANALYTICS
Don't Worry about Spilled Beans



K- Means Clustering Output



Labelled Data

K - Means Clustering: Model Evaluation

BEANALYTICS
Don't Worry about Spilled Beans

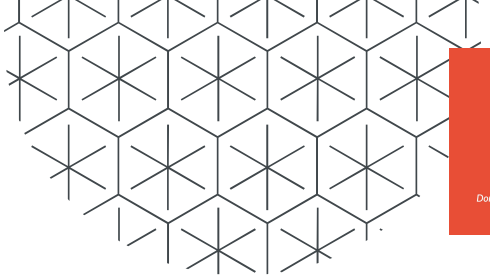
Model Evaluation Metric (self generated):

Our aim was to replicate the clustering pattern of the labels in the distribution of the data with the help of the Principal Components.

We wanted to see if the model was accurately clustering the beans into some distinct clusters and by comparing the positions of each individual points in the cluster with the original data, we managed to get an idea about the accuracy of the model in predicting certain class of beans.

Result of self generated accuracy scores:

```
The accuracy of BARBUNYA beans in our cluster analysis is 82.07%
The accuracy of BOMBAY beans in our cluster analysis is 99.62%
The accuracy of CALI beans in our cluster analysis is 83.01%
The accuracy of DERMASON beans in our cluster analysis is 83.08%
The accuracy of HOROZ beans in our cluster analysis is 85.89%
The accuracy of SEKER beans in our cluster analysis is 92.11%
The accuracy of SIRA beans in our cluster analysis is 85.77%
```

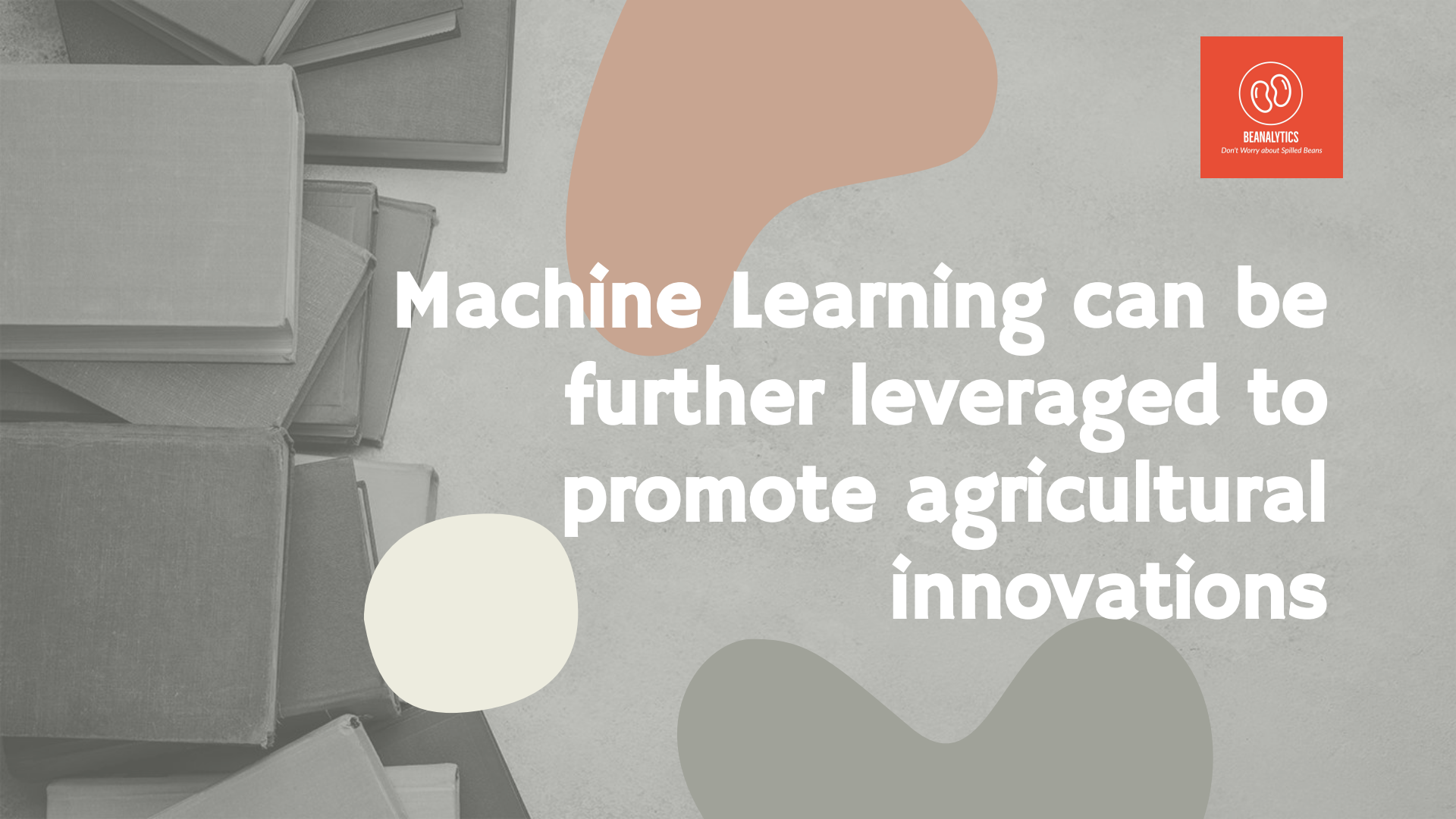


Conclusion

Beans are not trivial!

Summary of Results:

- Feature selection using PCA helps reduce data dimensionality and model complexity when data is highly correlated but important
- K-Means clustering and WCSS indicated that optimum number of clusters is 7 (similar to labelled data)
- Clusters with distinct attributes (eg Bombay) would result in higher accuracy in clustering than clusters/beans that have similar attributes
- There is scope for improving the classification process using train and test models using more complex classification algorithms such as random forests and XGboost.


BEANALYTICS
Don't Worry about Spilled Beans
Machine Learning can be further leveraged to promote agricultural innovations

Any Questions?