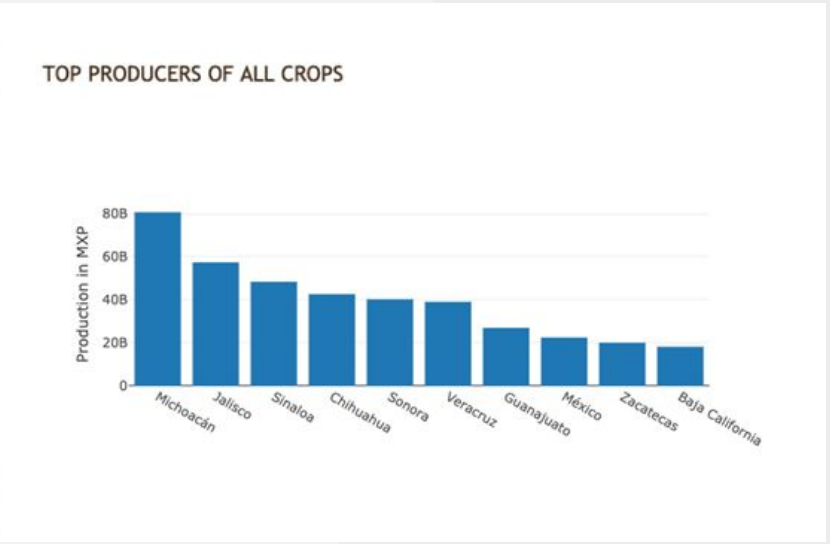
Agriculture in Mexico

Clustering Analysis of crops produced in Mexico in 2017

BACKGROUND



Upon completion of our last project, we realized out that **our web app could be improved to show deeper insights that includes other variables related to production volume** rather than just relying on those related to profitability

OBJECTIVE

The objective of this project is to use an unsupervised learning algorithm **KMeans Clustering** to analyze and discover new trends/patterns that are related to production volume across our dataset.

With this, we can **provide key insights that support decision making of mexican agriculture producers.** For performing such analysis, we make use of the “clean” dataset that we worked with during the past project.

Note: We gather this dataset from the national agricultural production of 2017 from Servicio de Información Agroalimentaria y Pesquera (SIAP), which is the agency in charge of generating statistics and geographic data on agriculture and agronomy.

SIAP
SERVICIO DE INFORMACIÓ
AGROALIMENTARIA Y PESQ

Clustering Walkthrough - Step 1: Importing, Exploring & Cleaning the Dataset

```
# Display Dataframe
df.head()
```

	anio	idest	estado	idmuni	municipio	ciclproductivo	modalidad	unidad	cultivo	sembrada	...	siniestrada	volumenproduccion	rendimiento
0	2017	16	Michoacán	75	Los Reyes	Perennes	Riego	Tonelada	Zarzamora	5088.0	...	0.0	131465.2	25.98
1	2017	16	Michoacán	83	Tancítaro	Perennes	Temporal	Tonelada	Aguacate	19502.0	...	0.0	179436.0	9.49
2	2017	16	Michoacán	108	Zamora	Otoño-Invierno	Riego	Tonelada	Fresa	3071.0	...	0.0	174952.2	56.97
3	2017	26	Sonora	17	Caborca	Perennes	Riego	Tonelada	Espárrago	8021.0	...	0.0	86755.2	10.82
4	2017	26	Sonora	30	Hermosillo	Perennes	Riego	Tonelada	Uva	9543.9	...	0.0	156183.0	17.57

5 rows × 21 columns

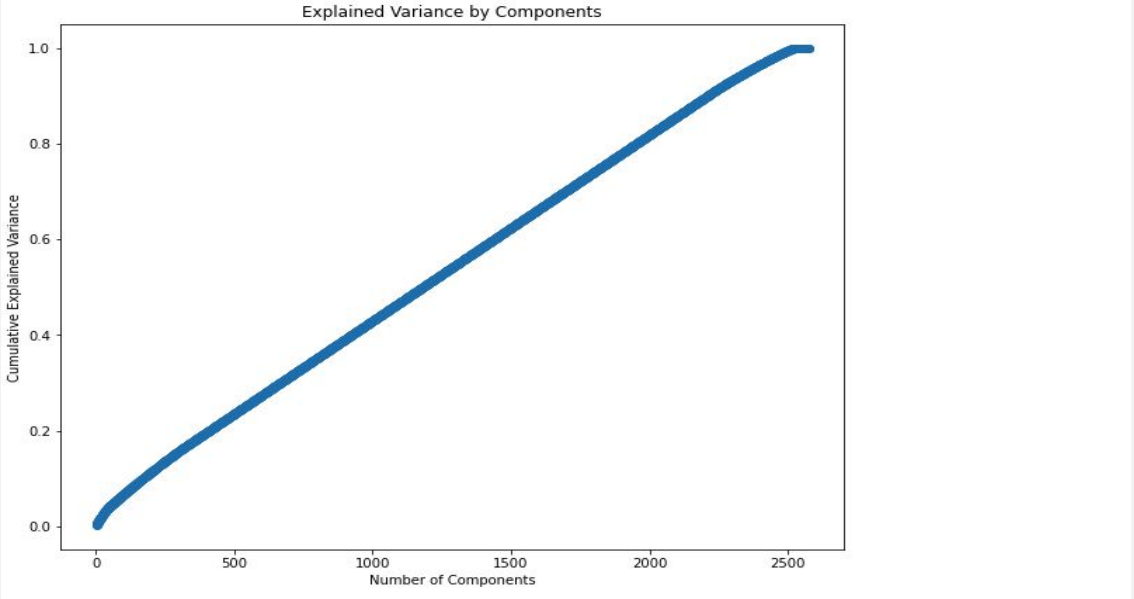
Key tasks we performed:

- **Drop 1493 null values** located on latitude, longitude & altitude columns.
- **Drop 11 columns** that we defined as **non-useful for this analysis**. Such as: anio, idest, estado, and some others.
  - We defined columns as **non-useful** when:
    i. By converting a categorical column into numeric our dataset significantly increase its **dimensions**. Example: Municipio, cultivo, estado columns.
    ii. The data does not add any value to the main objective of the clustering analysis. Example: precio, valorproduccion, **rendimiento** and so on.
- **Sort our dataset by the top 5 cultivos** that have more records, which are: **"Maíz Grano", "Frijol", "Avena Forrajera en verde", "Pastos y praderas" and "Tomate rojo (jitomate).**
  - There were more than 300 cultivos and dealing with all of them could cause a serious dimensionality issue. **You can see an example of this on the next slide.**

Example:

```
1  plt.figure(figsize=(10,8))
2  plt.plot(range(1,2576), pca.explained_variance_ratio_.cumsum(), marker='o', linestyle='--')
3  plt.title('Explained Variance by Components')
4  plt.xlabel('Number of Components')
5  plt.ylabel('Cumulative Explained Variance')
```

```
Text(0, 0.5, 'Cumulative Explained Variance')
```



Clustering Walkthrough - Step 2: Preprocessing Converting categorical features into numeric, Scaling & PCA

X.shape:
(11252, 10)

	cicloproductivo	modalidad	cultivo	sembrada	cosechada	siniestrada	volumenproduccion	latitud	longitud	altitud
20	Otoño-Invierno	Riego	Maíz grano	35012.00	35012.00	0.0	401061.32	24.767219	-107.696760	12.0
21	Primavera-Verano	Riego	Tomate rojo (jitomate)	1302.00	1302.00	0.0	108818.40	31.808944	-116.595134	18.0
28	Otoño-Invierno	Riego	Maíz grano	30597.34	30597.34	0.0	361643.68	25.783417	-108.994343	9.0
29	Otoño-Invierno	Riego	Maíz grano	30816.00	30816.00	0.0	364810.83	24.808808	-107.393756	57.0
31	Otoño-Invierno	Riego	Tomate rojo (jitomate)	2331.00	2331.00	0.0	232416.02	24.808808	-107.393756	57.0
...	...	...	...	...	...	...	...	...	...	...

1) `X = pd.get_dummies(X)`

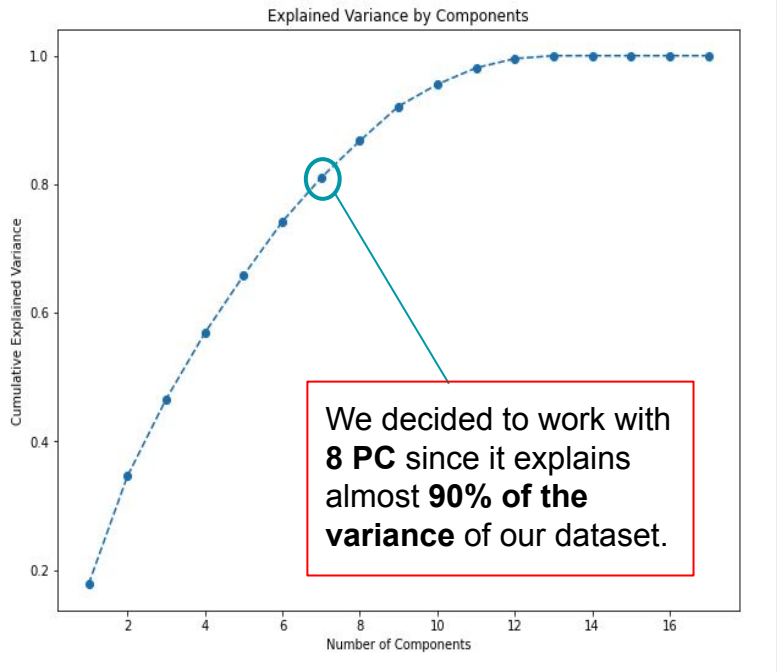
X.shape:
(11252, 17)

2) `from sklearn.preprocessing import StandardScaler`

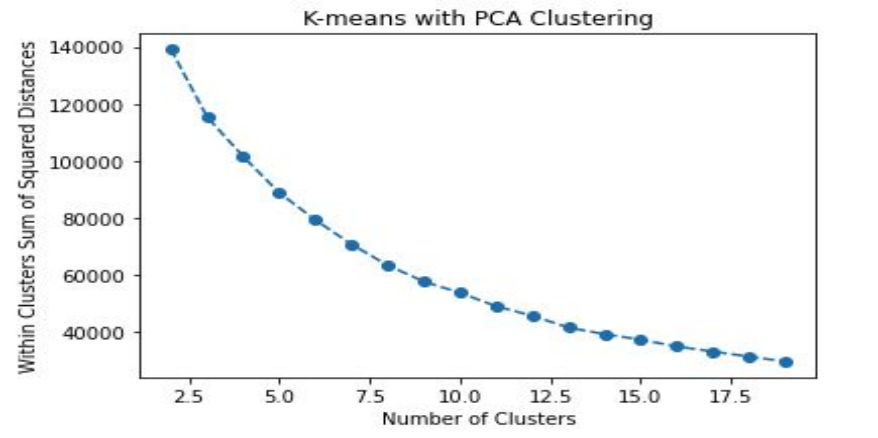
```
scaler = StandardScaler()
scaled_X = scaler.fit_transform(X)
```

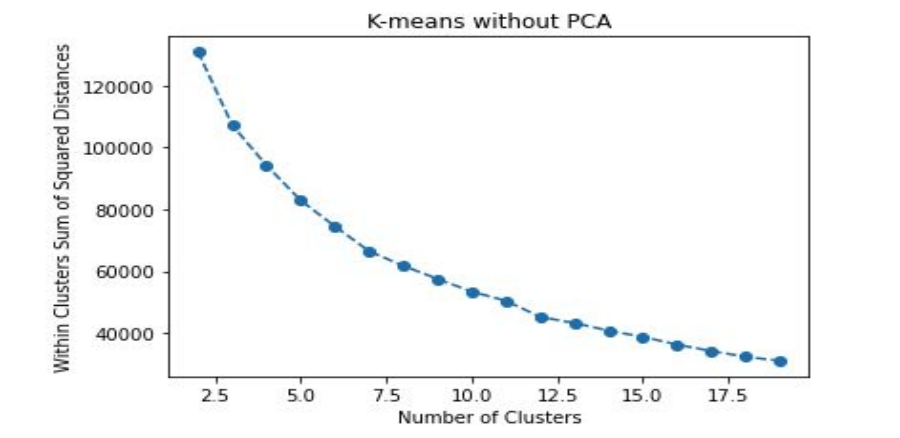
3) `from sklearn.decomposition import PCA`

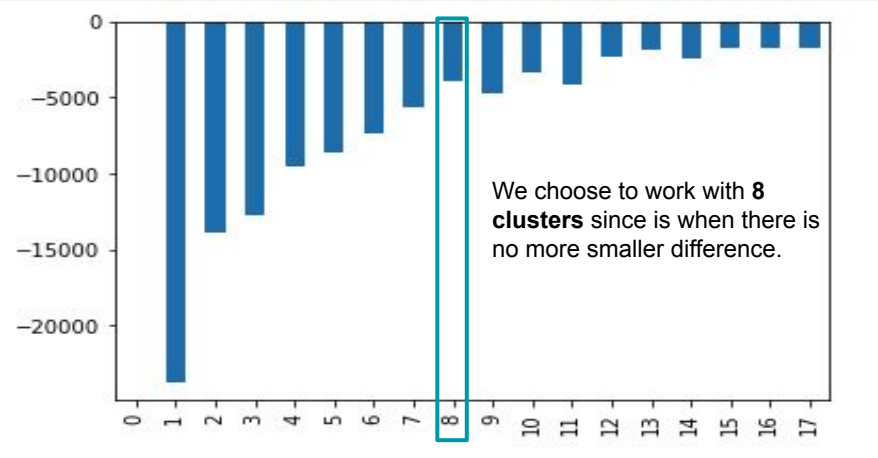
```
pca = PCA()
pca.fit_transform(scaled_X)
```



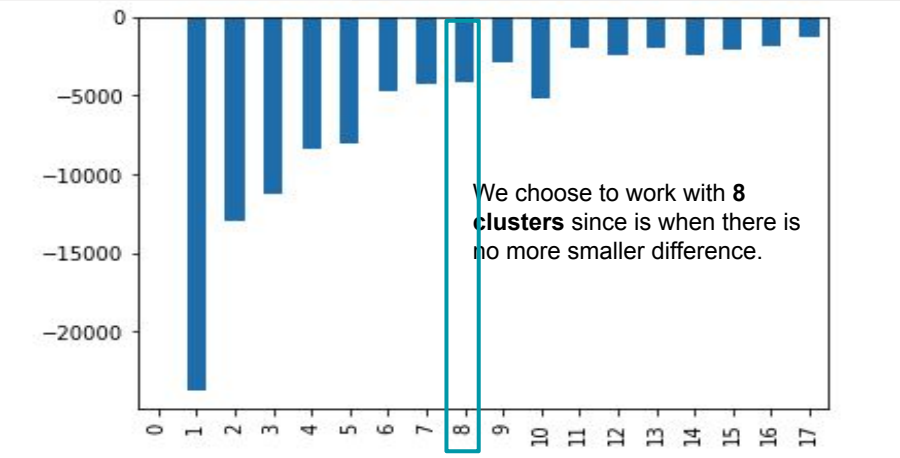
Clustering Walkthrough - Step 3: Implementing KMeans Algorithm







We choose to work with **8 clusters** since is when there is no more smaller difference.



We choose to work with **8 clusters** since is when there is no more smaller difference.

Clustering Walkthrough - Step 5: Is it worth it to combine PCA with K-means?

```
cluster_df.groupby(['K=8 Cluster Labels'])['rendimiento'].mean()
```

Results with PCA:

```
K=8 Cluster Labels
0       1.955123
1     103.885517
2      26.148438
3       0.873554
4      18.656641
5       2.311400
6       4.733247
7      20.416483
Name: rendimiento, dtype: float64
```

Results without PCA:

```
K=8 Cluster Labels
0      19.127143
1       1.554670
2       0.894091
3      26.156683
4     102.019901
5       5.965058
6       1.962511
7      15.955745
Name: rendimiento, dtype: float64
```

How do you calculate rendimiento?

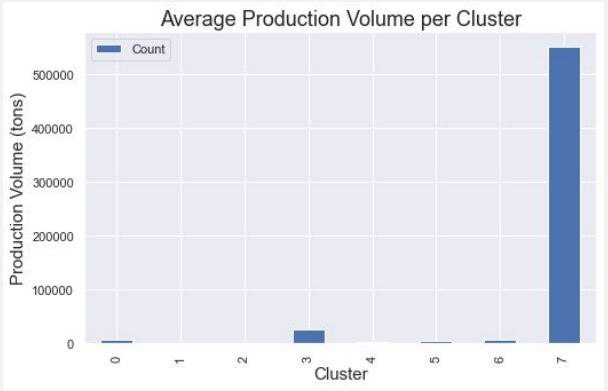
Volumenproduccion (tons) / cosechada (harvested area in hectares) = Rendimiento tons you get per harvested area in hectares

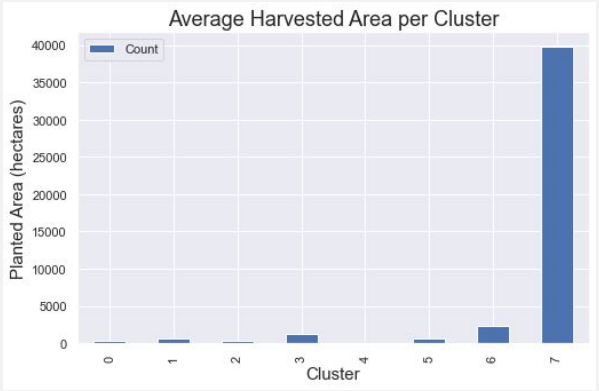
Clustering Visualizations

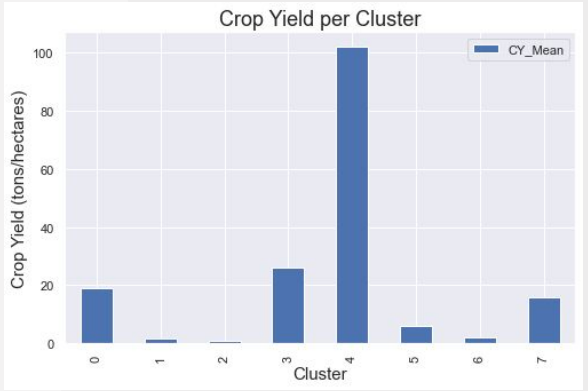
The algorithms led us to obtain 8 main clusters from the **8 different clusters with the 5 most cultivated products in Mexico**, with this data we were able to create a powerful visualization that enables the user analyze different patterns on rendimiento (tons you get by harvested area in hectares).

But why the clusters performance is so different? In order to answer this question we carry out a more in-depth analysis of the clusters.

UNDERSTANDING ‘CROP YIELD’ *(Rendimiento)*







Cluster No.	Production Volume
0	6,357
1	992
2	250
3	24,857
4	2,800
5	4,847
6	5,893
7	550,938

Cluster No.	Harvested Area
0	409
1	578
2	370
3	1,190
4	41
5	632
6	2,358
7	39,733

Cluster No.	Harvested Area
0	19.13
1	1.55
2	0.89
3	26.16
4	102.02
5	5.97
6	1.96
7	15.96

CROP YIELD = PRODUCTION VOLUME / HARVESTED HECTARES

(DOESN’T WORK WITH AVERAGES BECAUSE OF ROUNDING ISSUES)

INVESTOR'S ROLE

As an investor, we would expect to invest in a crop with the following characteristics:

- Reduced area: cheaper investment (considered)
- High production volume: larger volume to be sold (considered)
- Crop according to region: assures weather conditions (considered)
- Easy crop to grow: that have a low ratio of sinistered hectares (considered)
- Crop that sells the most expensive: greater income (not considered)
- Crop that has demand in the market: growth opportunity (not considered)
- Low investment in equipment: cheaper to start (not considered)
- Greatest revenue margin: therefore the best investment (not considered)

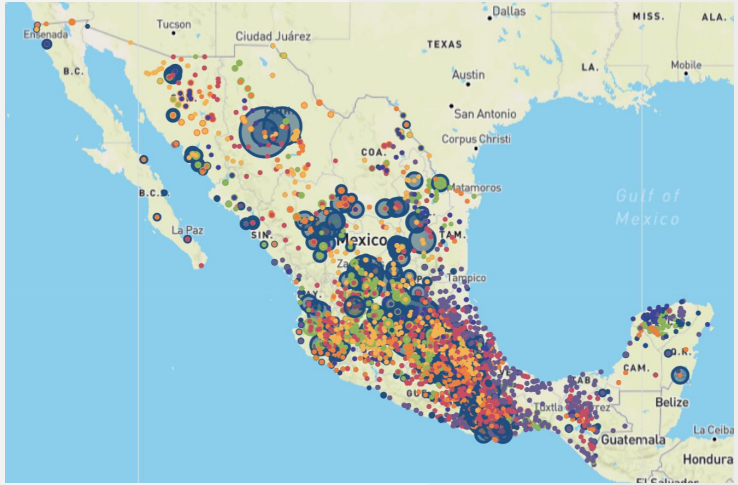
CLUSTER ANALYSIS

state	K = 0	K = 1	K = 2	K = 3	K = 4	K = 5	K = 6	K = 7	TOTAL
Oaxaca	37	209	558	96	352	601	552	6	2411
Puebla	104	59	195	8	78	174	209	0	827
Veracruz	14	270	105	10	21	44	204	0	668
Michoacán	152	1	73	66	100	149	115	1	657
México	186	0	88	83	74	80	116	0	627
Jalisco	99	12	67	151	54	97	106	8	594
Hidalgo	102	38	107	32	42	78	82	0	481
Guerrero	2	49	72	65	71	122	76	1	458
Yucatán	0	16	1	192	53	26	106	5	399
Chiapas	0	139	83	17	18	25	115	1	398
Zacatecas	117	0	78	54	32	54	55	7	397
San Luis Potosí	70	40	53	43	28	47	62	4	347
Tlaxcala	105	0	85	7	13	55	59	0	324
Sonora	70	1	40	23	29	115	17	0	295
Chihuahua	92	0	72	17	8	65	37	2	293
Durango	82	0	53	25	15	49	39	1	264
Guanajuato	46	0	62	16	39	47	47	0	257
Coahuila	64	2	20	52	10	53	23	0	224
Sinaloa	0	14	4	32	35	104	16	8	213
Tamaulipas	6	22	16	65	10	38	27	0	184
Morelos	6	0	29	13	44	49	29	0	170
Nuevo León	26	9	15	65	8	18	21	1	163
Nayarit	0	20	7	30	17	28	19	0	121
Querétaro	16	0	25	2	9	27	18	0	97
Aguascalientes	21	0	17	17	9	16	10	0	90
Tabasco	0	49	0	0	2	0	22	0	73
Colima	0	0	0	17	12	17	9	1	56
Campeche	0	20	2	0	7	10	14	1	54
Baja California	11	0	0	7	7	22	1	0	48
Quintana Roo	0	15	1	0	6	6	8	0	36
Baja California Sur	0	0	3	1	10	12	0	0	26

Total records per state per cluster

crop	K = 0	K = 1	K = 2	K = 3	K = 4	K = 5	K = 6	K = 7	TOTAL
Avena forrajera en verde	1428	0	0	0	0	0	0	1	1429
Frijol	0	449	1930	0	0	332	7	9	2727
Maíz grano	0	521	0	0	0	1743	2197	11	4472
Pastos y praderas	0	3	1	1206	0	153	10	26	1399
Tomate rojo (jitomate)	0	12	0	0	1213	0	0	0	1225

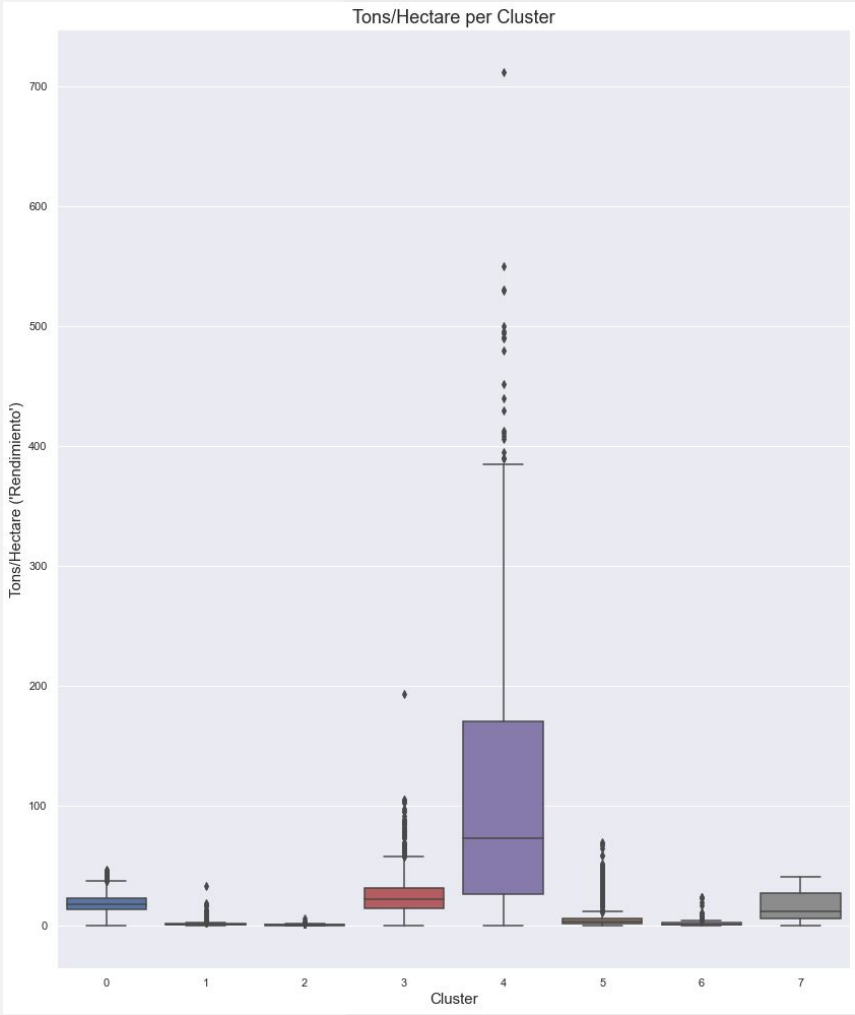
Total records per crop per cluster



The **boxplot** diagram shown in the image allows us to visualize the crop yield variable on each of the clusters. The tables below show the stats detail as well as the number and percentage of outliers per cluster.

Cluster No.	count	mean	std	min	25%	50%	75%	max	IQR
0	1428	19.1271	8.3267	0	13.46	18	23	47	9.54
1	985	1.5547	1.9631	0.04	0.78	1.1	1.72	33.6	0.94
2	1931	0.8941	0.5008	0	0.55	0.76	1.1	6.45	0.55
3	1206	26.1567	17.8308	0	14.4	22.37	31.83	193.45	17.43
4	1213	102.0199	91.8266	0	26.5	73.33	170.25	711.68	143.75
5	2228	5.9651	8.3379	0	2.29	3.45	6.1	69.8	3.81
6	2214	1.9625	1.8070	0	0.94	1.3	2.52	24	1.58
7	47	15.9557	13.2127	0.34	6.425	11.82	27.29	41.11	20.865

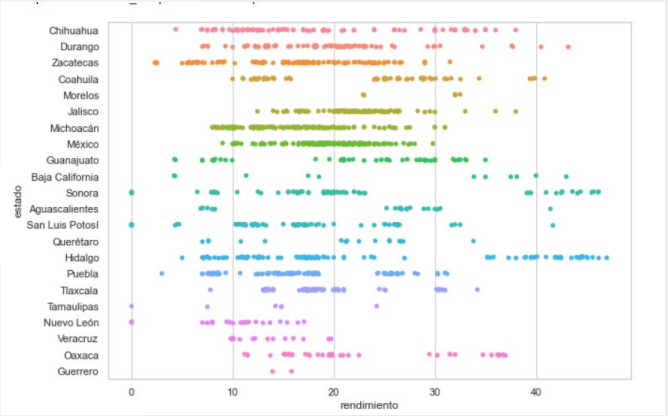
Cluster No.	Outlier count	Outlier Percentage
0	63	4.41%
1	71	7.21%
2	95	4.92%
3	70	5.80%
4	21	1.73%
5	167	7.50%
6	125	5.65%
7	0	0.00%



Cluster #0

- The cluster 0 only includes one crop, “Avena forrajera en verde”

	cultivo_Avena forrajera en verde	cultivo_Frijol	cultivo_Maíz grano	cultivo_Pastos y praderas	cultivo_Tomate rojo (jitomate)
K=8 Cluster Labels					
0	1	0	0	0	0



- The state with the biggest mean production is ‘Mexico’ and the lowest is ‘Guerrero’.

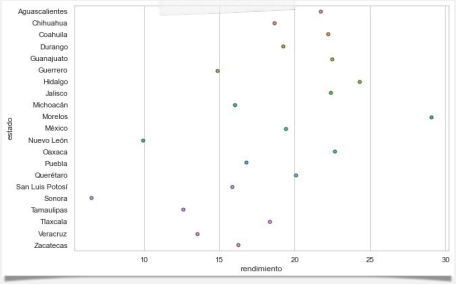
	cultivo_Avena forrajera en verde
estado	
Aguascalientes	21
Baja California	11
Chihuahua	92
Coahuila	64
Durango	82
Guanajuato	46
Guerrero	2
Hidalgo	102
Jalisco	99
Michoacán	152
Morelos	6
México	186
Nuevo León	26
Oaxaca	37
Puebla	104
Querétaro	16
San Luis Potosí	70
Sonora	70
Tamaulipas	6
Tlaxcala	105
Veracruz	14
Zacatecas	117

Cluster #1

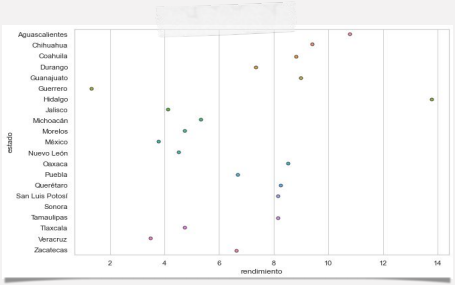
- For cluster 1 the main crop is “Maíz grano”

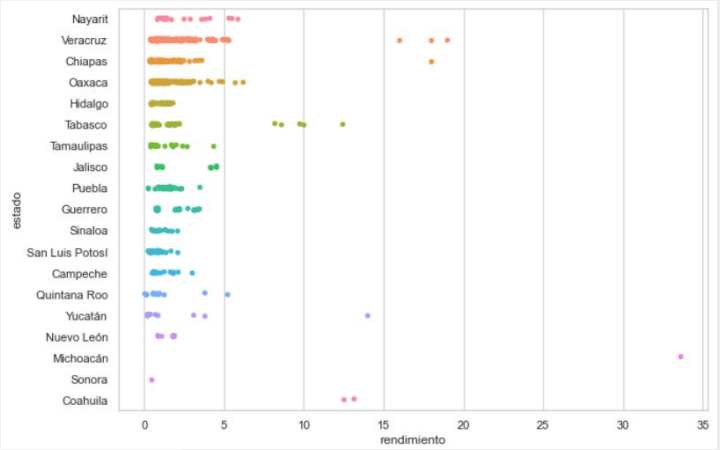
K=8 Cluster Labels	cultivo_Avena forrajera en verde	cultivo_Frijol	cultivo_Maiz grano	cultivo_Pastos y praderas	cultivo_Tomate rojo (jitomate)
1	0.0	0.455838	0.528934	0.003046	0.012183

- The state with the biggest mean production is ‘Michoacán’ and the lowest is ‘Sonora’.



- Once we analyze the standard deviation it becomes clear that the production is volatile so we can take as conclusions the results of the mean production by crop.



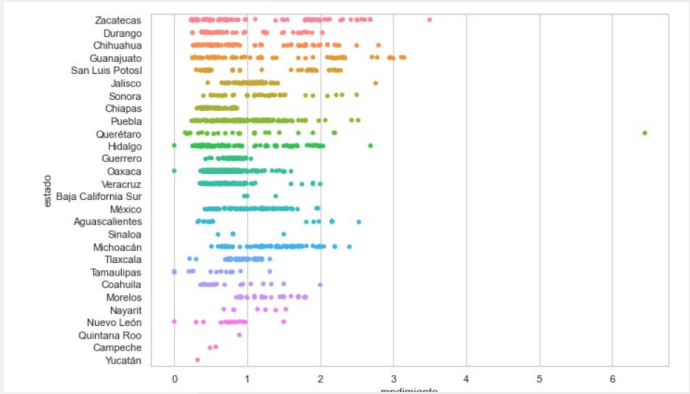


Cluster #2

The cluster 1 only includes two crops, “Frijol” and “Pastos y praderas”

	cultivo_Avena forrajera en verde	cultivo_Frijol	cultivo_Maíz grano	cultivo_Pastos y praderas	cultivo_Tomate rojo (jitomate)
K=8 Cluster Labels					
2	0.0	0.999482	0.0	0.000518	0.0

T



The state with the biggest mean production is ‘Oaxaca’ and the lowest is ‘Quintana Roo’.

The state with the biggest mean production is ‘Queratarо’ and the lowest are the same for the rest

estado	cultivo_Frijol
Aguascalientes	17
Baja California Sur	3
Campeche	2
Chiapas	83
Chihuahua	72
Coahuila	20
Durango	53
Guanajuato	62
Guerrero	72
Hidalgo	107
Jalisco	67
Michoacán	73
Morelos	29
México	88
Nayarit	7
Nuevo León	15
Oaxaca	558
Puebla	195
Querétaro	24
Quintana Roo	1
San Luis Potosí	53
Sinaloa	4
Sonora	40
Tamaulipas	16

estado	cultivo_Pastos y praderas
Aguascalientes	0
Baja California Sur	0
Campeche	0
Chiapas	0
Chihuahua	0
Coahuila	0
Durango	0
Guanajuato	0
Guerrero	0
Hidalgo	0
Jalisco	0
Michoacán	0
Morelos	0
México	0
Nayarit	0
Nuevo León	0
Oaxaca	0
Puebla	0
Querétaro	1
Quintana Roo	0
San Luis Potosí	0
Sinaloa	0
Sonora	0
Tamaulipas	0
Tlaxcala	0
Veracruz	0
Yucatán	0
Zacatecas	0

Cluster #3

General observations on this cluster:

- This cluster includes only data regarding the crop *“pastos y praderas”*. 86% of this crop’s data is contained in cluster no. 3.

Avena forrajera en verde	0
Frijol	0
Maíz grano	0
Pastos y praderas	1206
Tomate rojo (jitomate)	0

- Most of the considered crops (99%) are perennial (they don’t need to be replanted each year because after harvest they automatically grow back).

Ciclo	K = 0	K = 1	K = 2	K = 3	K = 4	K = 5	K = 6	K = 7
Otoño-Invierno	744	985	0	4	445	1171	0	8
Primavera-Verano	684	0	1931	2	768	1057	2214	13
Perennes	0	0	0	1200	0	0	0	26

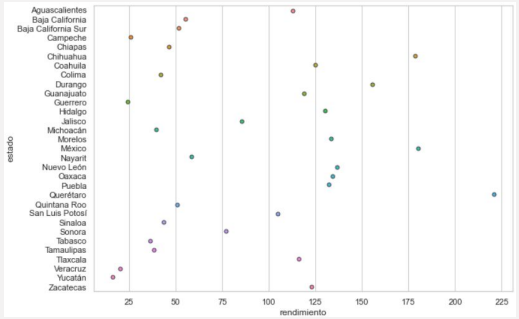
- The crop yield for the crops included in this cluster was 31,545 metric tons per hectare.

Cluster #4

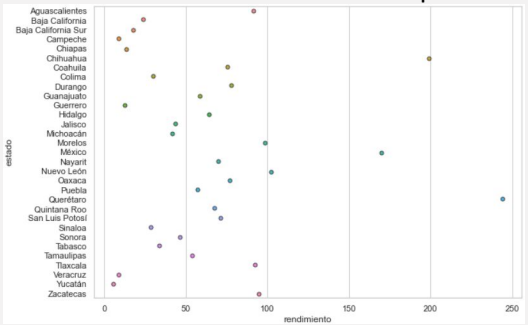
- The cluster 4 only includes one crop, “Tomate rojo (jitomate)”

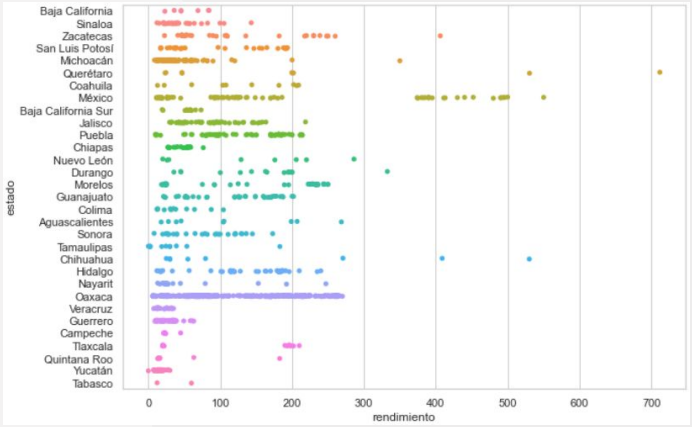
K=8 Cluster Labels	cultivo_Avena forrajera en verde	cultivo_Frijol	cultivo_Maíz grano	cultivo_Pastos y praderas	cultivo_Tomate rojo (jitomate)
4	0	0	0	0	1

- The state with the biggest mean production is ‘Querétaro’ and the lowest is ‘Yucatán’.



- Once we analyze the standard deviation it becomes clear that the production is volatile so we can take as conclusions the results of the mean production by crop.



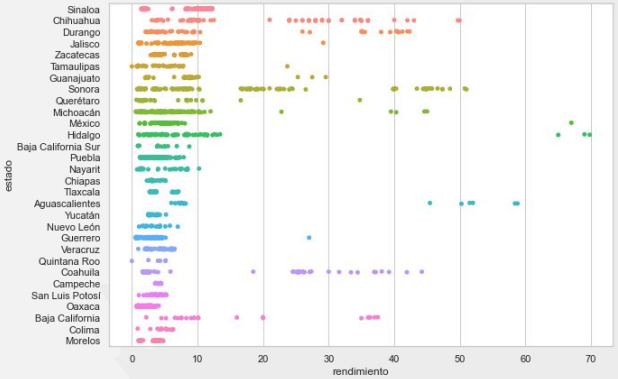


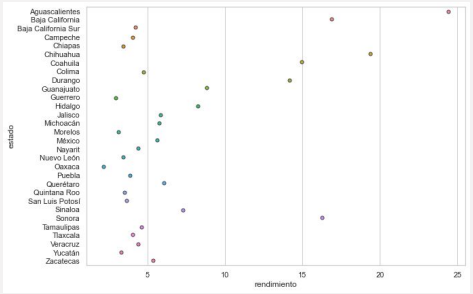
Cluster #5

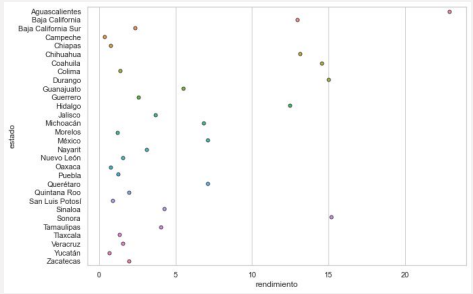
The cluster 5 includes “Frijol”, “Pastos y Praderas” and “Maíz grano”, being the “Maíz grano” the one with the highest average production within the cluster.

K=8 Cluster Labels	cultivo_Avena forrajera en verde	cultivo_Frijol	cultivo_Maíz grano	cultivo_Pastos y praderas	cultivo_Tomate rojo (jitomate)
5	0.0	0.149013	0.782316	0.068671	0.0

The state with the biggest mean production is ‘Aguascalientes’ and the lowest is ‘Oaxaca’. These doesn't change when we analyze the standard deviation by states.







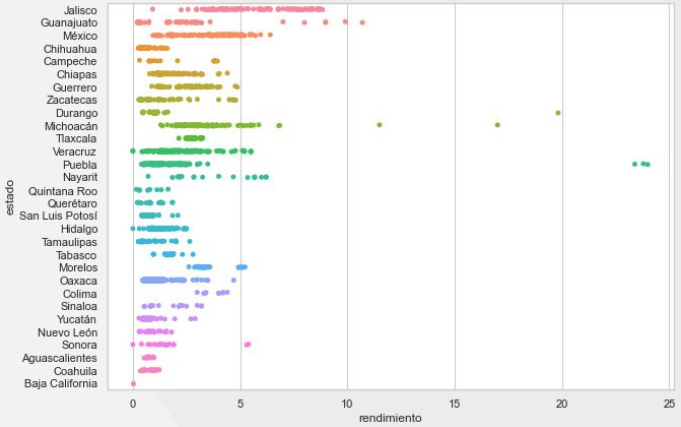
We can observe at the map that almost all the production is at the southern states.

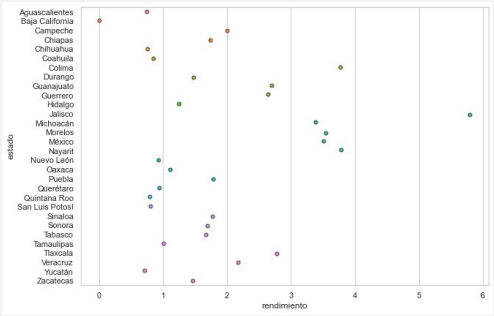
Cluster #6

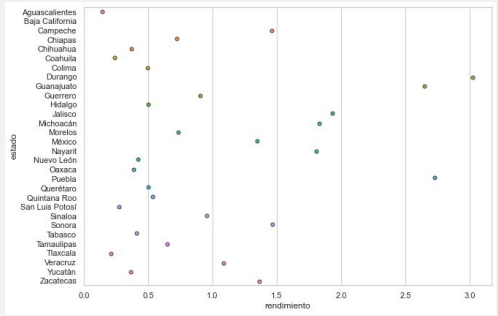
The cluster 6 includes “Frijol”, “Pastos y Praderas” and “Maíz grano”,
being the “Maíz grano” the one with the highest average production within the cluster.

K=8 Cluster Labels	cultivo_Avena forrajera en verde	cultivo_Frijol	cultivo_Maíz grano	cultivo_Pastos y praderas	cultivo_Tomate rojo (jitomate)
6	0.0	0.003162	0.992322	0.004517	0.0

The state with the biggest mean production is “Jalisco” and the lowest is “Baja California”.
However when we analyze the standard deviation by state we can observe that these changes.







We can observe at the map that almost all the production is at the southern states.

Cluster #7

General observations on this cluster:

- This cluster is the smallest one, containing only 47 data records, which is the 0.4% of all the data records in the data set.

Avena forrajera en verde	1
Frijol	9
Maíz grano	11
Pastos y praderas	26
Tomate rojo (jitomate)	0

- Most of the considered crops are perennial *(pastos y praderas)*, but it also includes data from all of the year seasons.

Ciclo	K = 0	K = 1	K = 2	K = 3	K = 4	K = 5	K = 6	K = 7
Otoño-Invierno	744	985	0	4	445	1171	0	8
Primavera-Verano	684	0	1931	2	768	1057	2214	13
Perennes	0	0	0	1200	0	0	0	26

- The crop yield for the crops included in this cluster was 750 metric tons per hectare (also the smallest from all clusters).

Bonus

	cicloproductivo	modalidad	cultivo	sembrada	cosechada	siniestrada	volumenproduccion	latitud	longitud	altitud
20	Otoño-Invierno	Riego	Maíz grano	35012.00	35012.00	0.0	401061.32	24.767219	-107.696760	12.0
21	Primavera-Verano	Riego	Tomate rojo (jitomate)	1302.00	1302.00	0.0	108818.40	31.808944	-116.595134	18.0
28	Otoño-Invierno	Riego	Maíz grano	30597.34	30597.34	0.0	361643.68	25.783417	-108.994343	9.0
29	Otoño-Invierno	Riego	Maíz grano	30816.00	30816.00	0.0	364810.83	24.808808	-107.393756	57.0
31	Otoño-Invierno	Riego	Tomate rojo (jitomate)	2331.00	2331.00	0.0	232416.02	24.808808	-107.393756	57.0
...	...	...	...	...	...	...	...	...	...	...

We perform clustering with this new DF...

	cicloproductivo	modalidad	cultivo	sembrada	latitud	longitud	altitud
20	Otoño-Invierno	Riego	Maíz grano	35012.00	24.767219	-107.696760	12.0
21	Primavera-Verano	Riego	Tomate rojo (jitomate)	1302.00	31.808944	-116.595134	18.0
28	Otoño-Invierno	Riego	Maíz grano	30597.34	25.783417	-108.994343	9.0
29	Otoño-Invierno	Riego	Maíz grano	30816.00	24.808808	-107.393756	57.0
31	Otoño-Invierno	Riego	Tomate rojo (jitomate)	2331.00	24.808808	-107.393756	57.0
...	...	...	...	...	...	...	...

Bonus

Note: we pass categorical inputs as number because we train the cluster algo with a df with dummies.

In [57]:
```
# Create the input data according to the dataframe structure you pass to your clustering3
input_data = [[20000, 29.0892,-110.961,202,1,0,0,1,0,0,1,0,0,0]]
```

sembrada, latitud, longitud, altitud, Otoño-Invierno, Perennes, Primavera-Verano, Riego, Temporal, Frijol, Maíz, Pastos y praderas, Tomate

In [58]:
```
# Scaled the input data
scaled_input = scaler.transform(input_data)
```

In [59]:
```
# Pass the scaled data into the model and save the results
result = model.transform(scaled_input)
```

In [60]:
```
# Check the output of the model
print(result)
```

```
[[8.09881262 7.63777995 8.22133019 7.55123188 7.56373858 6.85039246
  7.96842085]]
```

In [61]:
```
# Get the index of the minimum value of the results
np.argmin(result)
```

Out[61]: 5

In [62]:
```
# How does the cluster 5 looks like?
cluster_5_df = cluster_df[cluster_df['K=7 Cluster Labels'] == 5]
```

In [68]:
```
# Look for the same crop you selected in your input data inside cluster 5
cluster_5_df_frijol = cluster_5_df[cluster_5_df['cultivo']=='Frijol']
```

In [72]:
```
# Get the descriptive stats of crop yield for that specific crop in cluster 5
cluster_5_df_frijol['rendimiento'].describe()
```

Out[72]:
```
count    780.000000
mean       0.961962
std        0.448292
min        0.000000
25%        0.687500
50%        0.880000
75%        1.110000
max        2.980000
Name: rendimiento, dtype: float64
```

```
K=7 Cluster Labels
0      1.959738
1      5.049991
2     26.129276
3      0.890929
4     19.122239
5      4.397196
6    101.139608
Name: rendimiento, dtype: float64
```

Crop Yield Avg by Cluster. We get this from Clustering Model without the input data.