```
! pip install dataprep
```

```
Collecting dataprep
  Downloading dataprep-0.4.5-py3-none-any.whl.metadata (14 kB)
Requirement already satisfied: aiohttp<4.0,>=3.6 in /usr/local/lib/python3.10/dist-packages (from dataprep) (3.10.5)
Collecting bokeh<3,>=2 (from dataprep)
  Downloading bokeh-2.4.3-py3-none-any.whl.metadata (14 kB)
Requirement already satisfied: dask>=2022.3.0 in /usr/local/lib/python3.10/dist-packages (from dask[array,dataframe,delayed]>=2022.3.
Requirement already satisfied: flask<3,>=2 in /usr/local/lib/python3.10/dist-packages (from dataprep) (2.2.5)
Collecting flask_cors<4.0.0,>=3.0.10 (from dataprep)
  Downloading Flask_Cors-3.0.10-py2.py3-none-any.whl.metadata (5.4 kB)
Requirement already satisfied: ipywidgets<8.0,>=7.5 in /usr/local/lib/python3.10/dist-packages (from dataprep) (7.7.1)
Collecting jinja2<3.1,>=3.0 (from dataprep)
  Downloading Jinja2-3.0.3-py3-none-any.whl.metadata (3.5 kB)
Collecting jsonpath-ng<2.0,>=1.5 (from dataprep)
  Downloading jsonpath_ng-1.6.1-py3-none-any.whl.metadata (18 kB)
Collecting metaphone<0.7,>=0.6 (from dataprep)
  Downloading Metaphone-0.6.tar.gz (14 kB)
  Preparing metadata (setup.py) ... done
Requirement already satisfied: nltk<4.0.0,>=3.6.7 in /usr/local/lib/python3.10/dist-packages (from dataprep) (3.8.1)
Requirement already satisfied: numpy<2.0,>=1.21 in /usr/local/lib/python3.10/dist-packages (from dataprep) (1.26.4)
Collecting pandas<2.0,>=1.1 (from dataprep)
  Downloading pandas-1.5.3-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (11 kB)
Collecting pydantic<2.0,>=1.6 (from dataprep)
  Downloading pydantic-1.10.18-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (152 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 152.2/152.2 kB 4.9 MB/s eta 0:00:00
Requirement already satisfied: pydot<2.0.0,>=1.4.2 in /usr/local/lib/python3.10/dist-packages (from dataprep) (1.4.2)
Collecting python-crfsuite==0.9.8 (from dataprep)
  Downloading python_crfsuite-0.9.8-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (4.4 kB)
Collecting python-stdnum<2.0,>=1.16 (from dataprep)
  Downloading python_stdnum-1.20-py2.py3-none-any.whl.metadata (18 kB)
Collecting rapidfuzz<3.0.0,>=2.1.2 (from dataprep)
  Downloading rapidfuzz-2.15.2-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (9.0 kB)
Collecting regex<2022.0.0,>=2021.8.3 (from dataprep)
  Downloading regex-2021.11.10-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (39 kB)
Requirement already satisfied: scipy<2.0,>=1.8 in /usr/local/lib/python3.10/dist-packages (from dataprep) (1.13.1)
Collecting sqlalchemy==1.3.24 (from dataprep)
  Downloading SQLAlchemy-1.3.24.tar.gz (6.4 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 6.4/6.4 MB 59.0 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: tqdm<5.0,>=4.48 in /usr/local/lib/python3.10/dist-packages (from dataprep) (4.66.5)
Collecting varname<0.9.0,>=0.8.1 (from dataprep)
  Downloading varname-0.8.3-py3-none-any.whl.metadata (11 kB)
Requirement already satisfied: wordcloud<2.0,>=1.8 in /usr/local/lib/python3.10/dist-packages (from dataprep) (1.9.3)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0,>=3.6->dataprep)
```

```
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0,>=3.6->dataprep) (1.3.1)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0,>=3.6->dataprep) (24.2.0)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0,>=3.6->dataprep) (1.4.1
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0,>=3.6->dataprep) (6.0
Requirement already satisfied: yarl<2.0,>=1.0 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0,>=3.6->dataprep) (1.9.4)
Requirement already satisfied: async-timeout<5.0,>=4.0 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0,>=3.6->dataprep)
Requirement already satisfied: packaging>=16.8 in /usr/local/lib/python3.10/dist-packages (from bokeh<3,>=2->dataprep) (24.1)
Requirement already satisfied: pillow>=7.1.0 in /usr/local/lib/python3.10/dist-packages (from bokeh<3,>=2->dataprep) (9.4.0)
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib/python3.10/dist-packages (from bokeh<3,>=2->dataprep) (6.0.2)
Requirement already satisfied: tornado>=5.1 in /usr/local/lib/python3.10/dist-packages (from bokeh<3,>=2->dataprep) (6.3.3)
Requirement already satisfied: typing-extensions>=3.10.0 in /usr/local/lib/python3.10/dist-packages (from bokeh<3,>=2->dataprep) (4.1
Requirement already satisfied: click>=8.1 in /usr/local/lib/python3.10/dist-packages (from dask>=2022.3.0->dask[array,dataframe,delay
Requirement already satisfied: cloudpickle>=1.5.0 in /usr/local/lib/python3.10/dist-packages (from dask>=2022.3.0->dask[array,datafra
Requirement already satisfied: fsspec>=2021.09.0 in /usr/local/lib/python3.10/dist-packages (from dask>=2022.3.0->dask[array,datafram
Requirement already satisfied: partd>=1.4.0 in /usr/local/lib/python3.10/dist-packages (from dask>=2022.3.0->dask[array,dataframe,del
```

```
!pip install kaggle
```

```
Requirement already satisfied: kaggle in /usr/local/lib/python3.10/dist-packages (1.6.17)
Requirement already satisfied: six>=1.10 in /usr/local/lib/python3.10/dist-packages (from kaggle) (1.16.0)
Requirement already satisfied: certifi>=2023.7.22 in /usr/local/lib/python3.10/dist-packages (from kaggle) (2024.8.30)
Requirement already satisfied: python-dateutil in /usr/local/lib/python3.10/dist-packages (from kaggle) (2.8.2)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from kaggle) (2.32.3)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from kaggle) (4.66.5)
Requirement already satisfied: python-slugify in /usr/local/lib/python3.10/dist-packages (from kaggle) (8.0.4)
Requirement already satisfied: urllib3 in /usr/local/lib/python3.10/dist-packages (from kaggle) (2.0.7)
Requirement already satisfied: bleach in /usr/local/lib/python3.10/dist-packages (from kaggle) (6.1.0)
Requirement already satisfied: webencodings in /usr/local/lib/python3.10/dist-packages (from bleach->kaggle) (0.5.1)
Requirement already satisfied: text-unidecode>=1.3 in /usr/local/lib/python3.10/dist-packages (from python-slugify->kaggle) (1.3)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->kaggle) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->kaggle) (3.8)
```

```
!mkdir ~/.kaggle
```

```
!cp kaggle.json ~/.kaggle
```

```
!kaggle datasets download -d aksahaha/crop-recommendation
```

```
Warning: Your Kaggle API key is readable by other users on this system! To fix this, you can run 'chmod 600 /root/.kaggle/kaggle.json
Dataset URL: https://www.kaggle.com/datasets/aksahaha/crop-recommendation
License(s): copyright-authors
Downloading crop-recommendation.zip to /content
  0% 0.00/62.6k [00:00<?, ?B/s]
```

```
100% 62.6k/62.6k [00:00<00:00, 805kB/s]
```

```
!unzip -qq crop-recommendation.zip
```

LIBRARIES

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from plotly.subplots import make_subplots
import warnings
```

```
dt = pd.read_csv("/content/Crop_recommendation.csv")
```

```
dt.head()
```

	Nitrogen	phosphorus	potassium	temperature	humidity	ph	rainfall	label	Unnamed: 8	Unnamed: 9
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice	NaN	NaN
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice	NaN	NaN
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice	NaN	NaN
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice	NaN	NaN
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice	NaN	NaN

PRE-PROCESSING

```
dt.shape
```

```
(2200, 10)
```

```
dt.columns
```

```
Index(['Nitrogen', 'phosphorus', 'potassium', 'temperature', 'humidity', 'ph',
       'rainfall', 'label', 'Unnamed: 8', 'Unnamed: 9'],
      dtype='object')
```

```
dt.isnull().any()
```

	0
Nitrogen	False
phosphorus	False
potassium	False
temperature	False
humidity	False
ph	False
rainfall	False
label	False
Unnamed: 8	True
Unnamed: 9	True

dtype: bool

```
dt.isnull().sum()
```

	0
Nitrogen	0
phosphorus	0
potassium	0
temperature	0
humidity	0
ph	0
rainfall	0
label	0
Unnamed: 8	2200
Unnamed: 9	2200

dtype: int64

```
dt['label'].value_counts()
```

	label
rice	100
maize	100
jute	100
cotton	100
coconut	100
papaya	100
orange	100
apple	100
muskmelon	100
watermelon	100
grapes	100
mango	100
banana	100
pomegranate	100
lentil	100
blackgram	100
mungbean	100
mothbeans	100
pigeonpeas	100
kidneybeans	100
chickpea	100
coffee	100

dtype: int64

```
crop_summary = pd.pivot_table(dt,index=['label'],aggfunc='mean')
crop_summary.head()
```

label	Nitrogen	humidity	ph	phosphorus	potassium	rainfall	temperature
apple	20.80	92.333383	5.929663	134.22	199.89	112.654779	22.630942
banana	100.23	80.358123	5.983893	82.01	50.05	104.626980	27.376798
blackgram	40.02	65.118426	7.133952	67.47	19.24	67.884151	29.973340
chickpea	40.09	16.860439	7.336957	67.79	79.92	80.058977	18.872847
coconut	21.98	94.844272	5.976562	16.93	30.59	175.686646	27.409892

```
crop = dt.drop(['Unnamed: 8','Unnamed: 9'],axis=1)
```

```
crop.head()
```

	Nitrogen	phosphorus	potassium	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice

```
from dataprep.datasets import load_dataset
from dataprep.eda import create_report
create_report(crop)
```

```
Computing sum-0b8ee1f013a1a13a0758004c001b9f35:  72%|███████   | 1052/1461 [00:00<00:00, 2090.67it/s]          /usr/local/lib/python3
  return func(*(_execute_task(a, cache) for a in args))
/usr/local/lib/python3.10/dist-packages/dataprep/eda/distribution/render.py:274: FutureWarning: The frame.append method is deprecated
  df = df.append(pd.DataFrame({col: [nrows - npresent]}, index=["Others"]))
```

DataPrep Report | Overview | Variables ≡ | Interactions | Correlations | Missing Values

Overview

Dataset Statistics

Number of Variables	8
Number of Rows	2200
Missing Cells	0
Missing Cells (%)	0.0%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	258.2 KB
Average Row Size in Memory	120.2 B
Variable Types	Numerical: 7 Categorical: 1

Dataset Insights

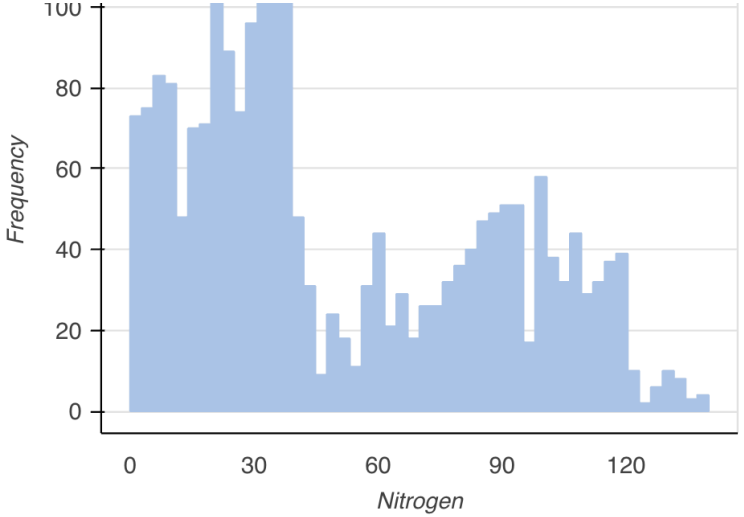
`potassium` is skewed	Skewed

Variables

Sort by Feature order ⌄ ☐ Reverse order

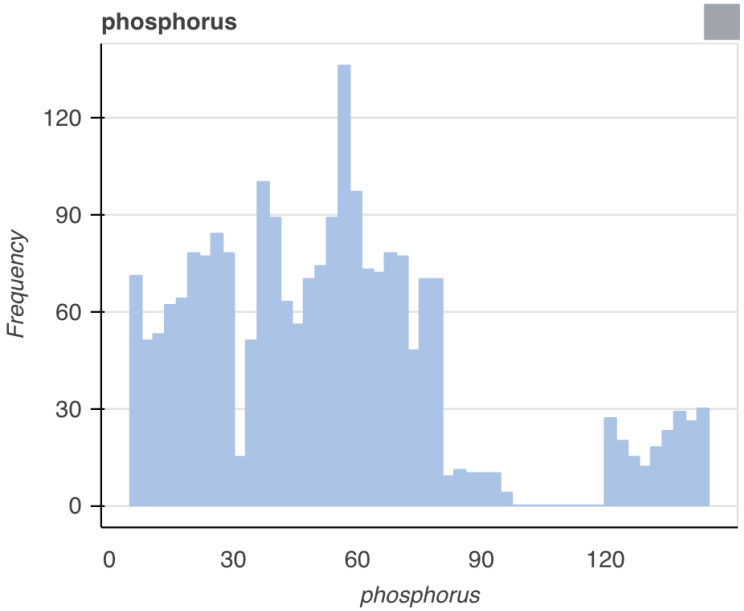
Nitrogen
numerical
Show Details

Approximate Distinct Count	137	Mean	50.5518
Approximate Unique (%)	6.2%	Minimum	0
Missing	0	Maximum	140
Missing (%)	0.0%	Zeros	27
Infinite	0	Zeros (%)	1.2%
Infinite (%)	0.0%	Negatives	0
Memory Size	35200	Negatives (%)	0.0%



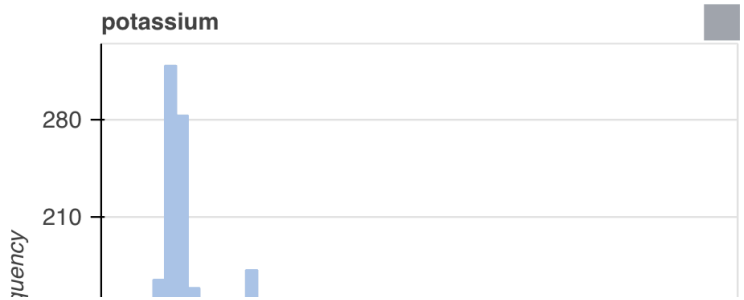
phosphorus
numerical
Show Details

Approximate Distinct Count	117	Mean	53.3627
Approximate Unique (%)	5.3%	Minimum	5
Missing	0	Maximum	145
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Negatives	0
Memory Size	35200	Negatives (%)	0.0%



potassium

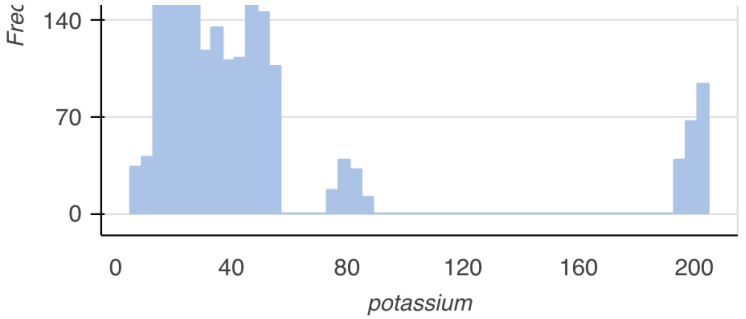
Approximate Distinct Count	73	Mean	48.1491
Approximate Unique (%)	3.3%	Minimum	5
Missing	0	Maximum	205



numerical

Show Details

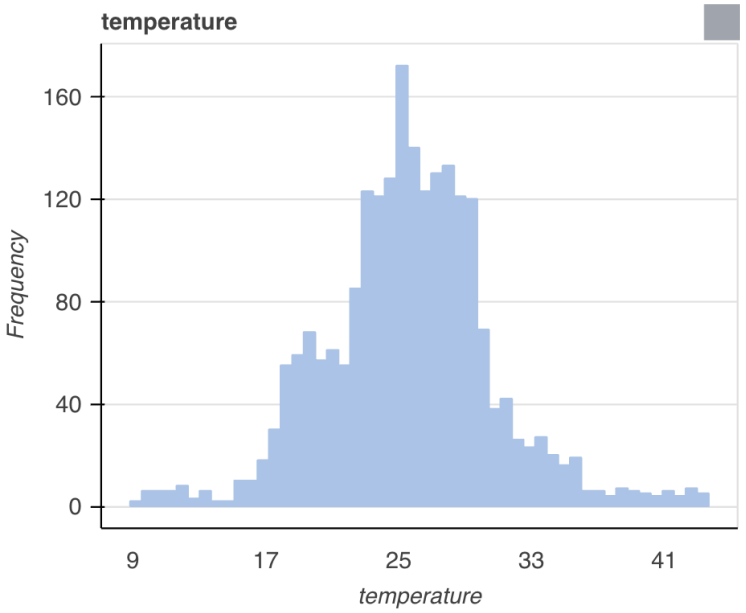
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Negatives	0
Memory Size	35200	Negatives (%)	0.0%



temperature
numerical

Show Details

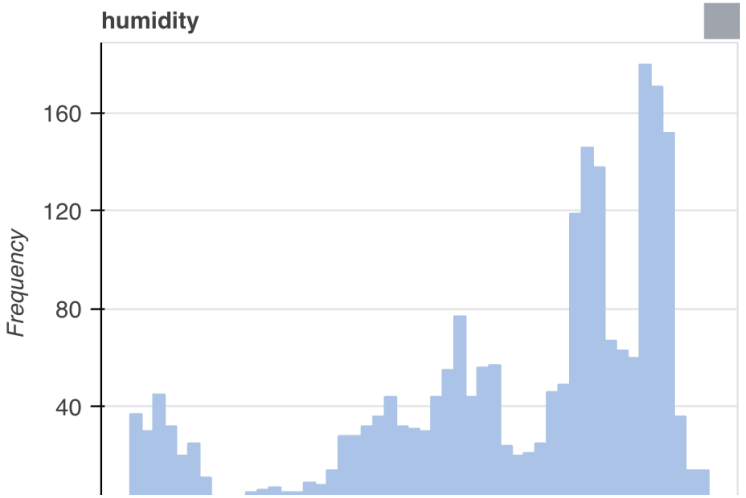
Approximate Distinct Count	2200	Mean	25.6162
Approximate Unique (%)	100.0%	Minimum	8.8257
Missing	0	Maximum	43.6755
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Negatives	0
Memory Size	35200	Negatives (%)	0.0%

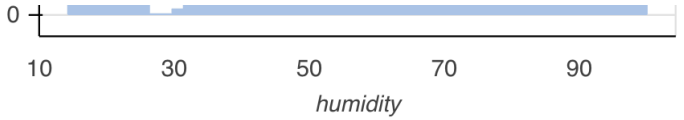


humidity
numerical

Show Details

Approximate Distinct Count	2200	Mean	71.4818
Approximate Unique (%)	100.0%	Minimum	14.258
Missing	0	Maximum	99.9819
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Negatives	0
Memory Size	35200	Negatives (%)	0.0%

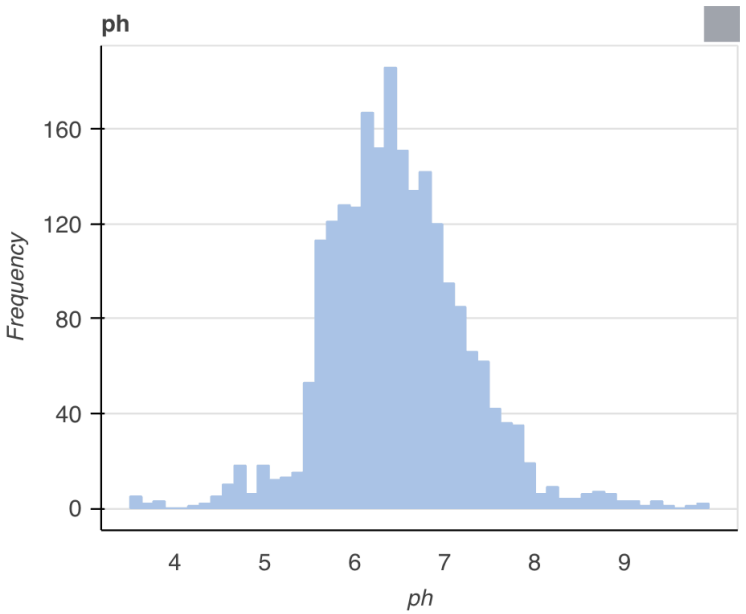




ph
numerical

Show Details

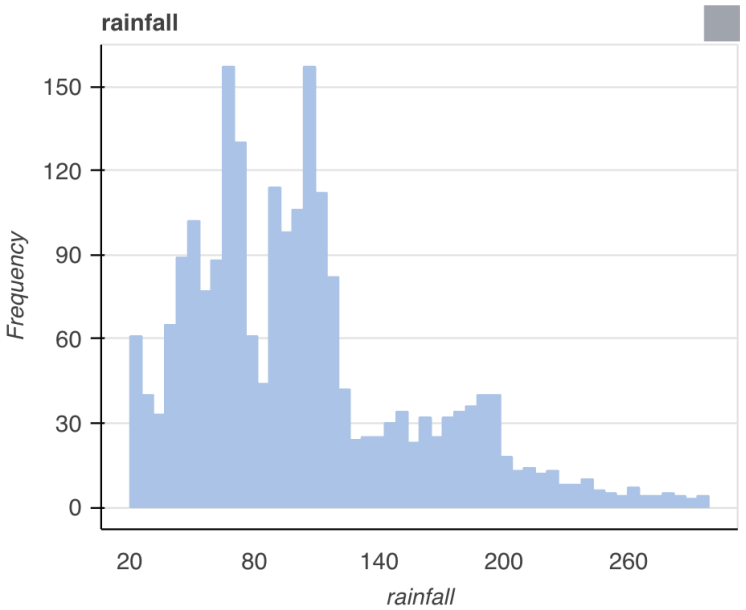
Approximate Distinct Count	2200	Mean	6.4695
Approximate Unique (%)	100.0%	Minimum	3.5048
Missing	0	Maximum	9.9351
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Negatives	0
Memory Size	35200	Negatives (%)	0.0%



rainfall
numerical

Show Details

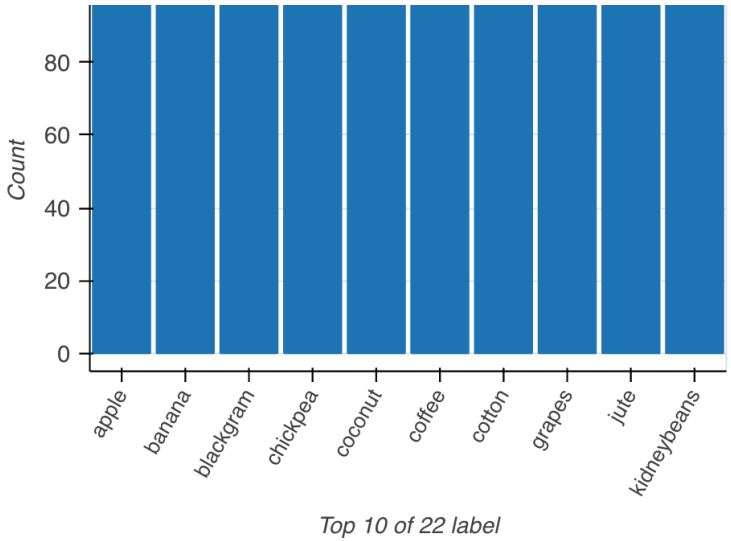
Approximate Distinct Count	2200	Mean	103.4637
Approximate Unique (%)	100.0%	Minimum	20.2113
Missing	0	Maximum	298.5601
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Negatives	0
Memory Size	35200	Negatives (%)	0.0%



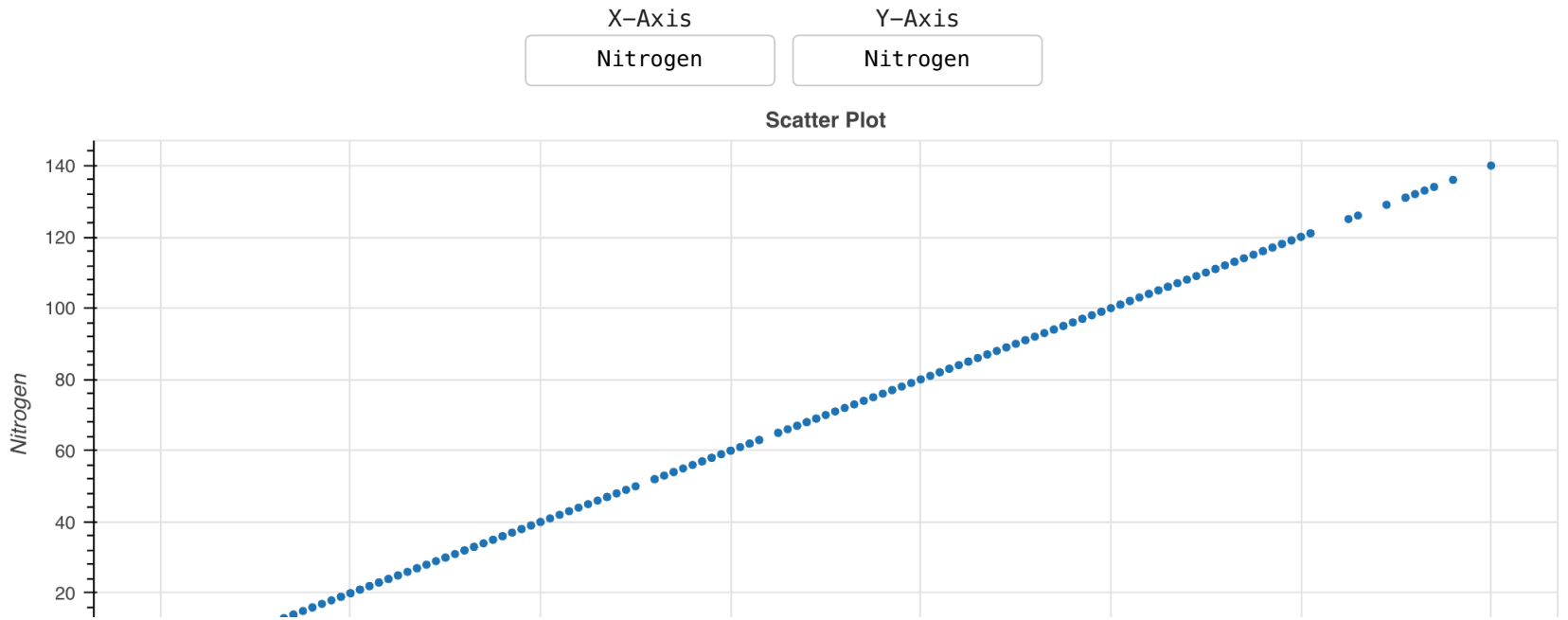
label
categorical

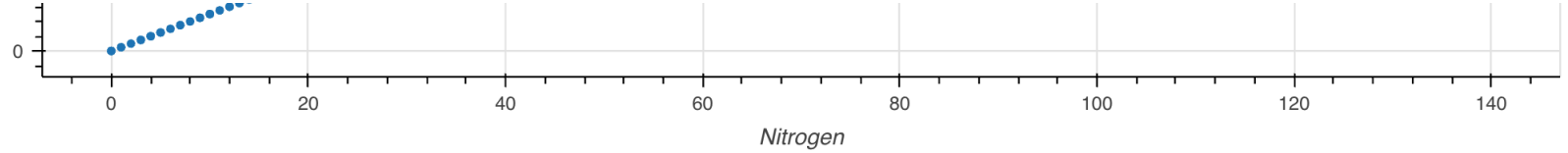
Show Details

Approximate Distinct Count	**22**
Approximate Unique (%)	**1.0%**
Missing	**0**
Missing (%)	**0.0%**
Memory Size	**158700**

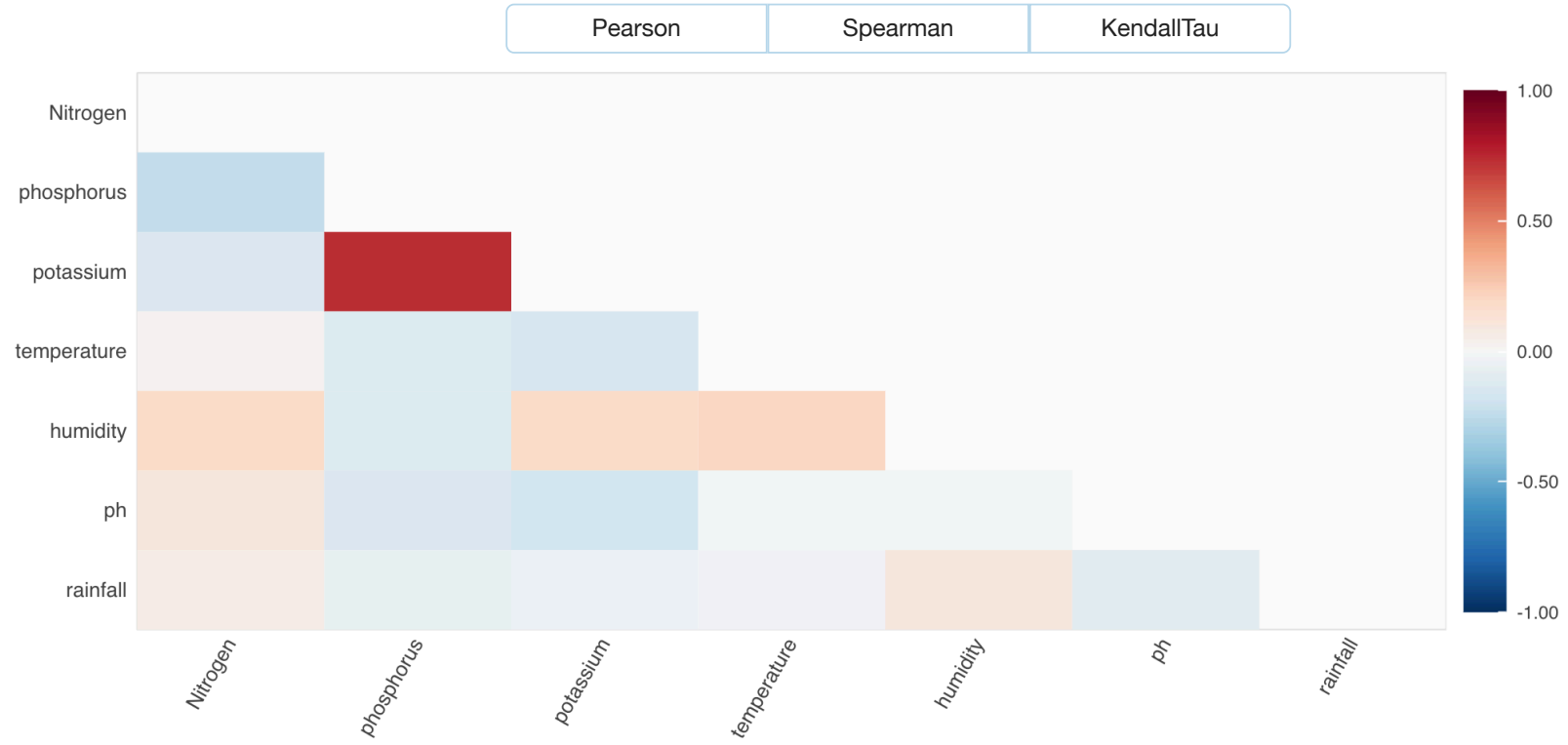


Interactions

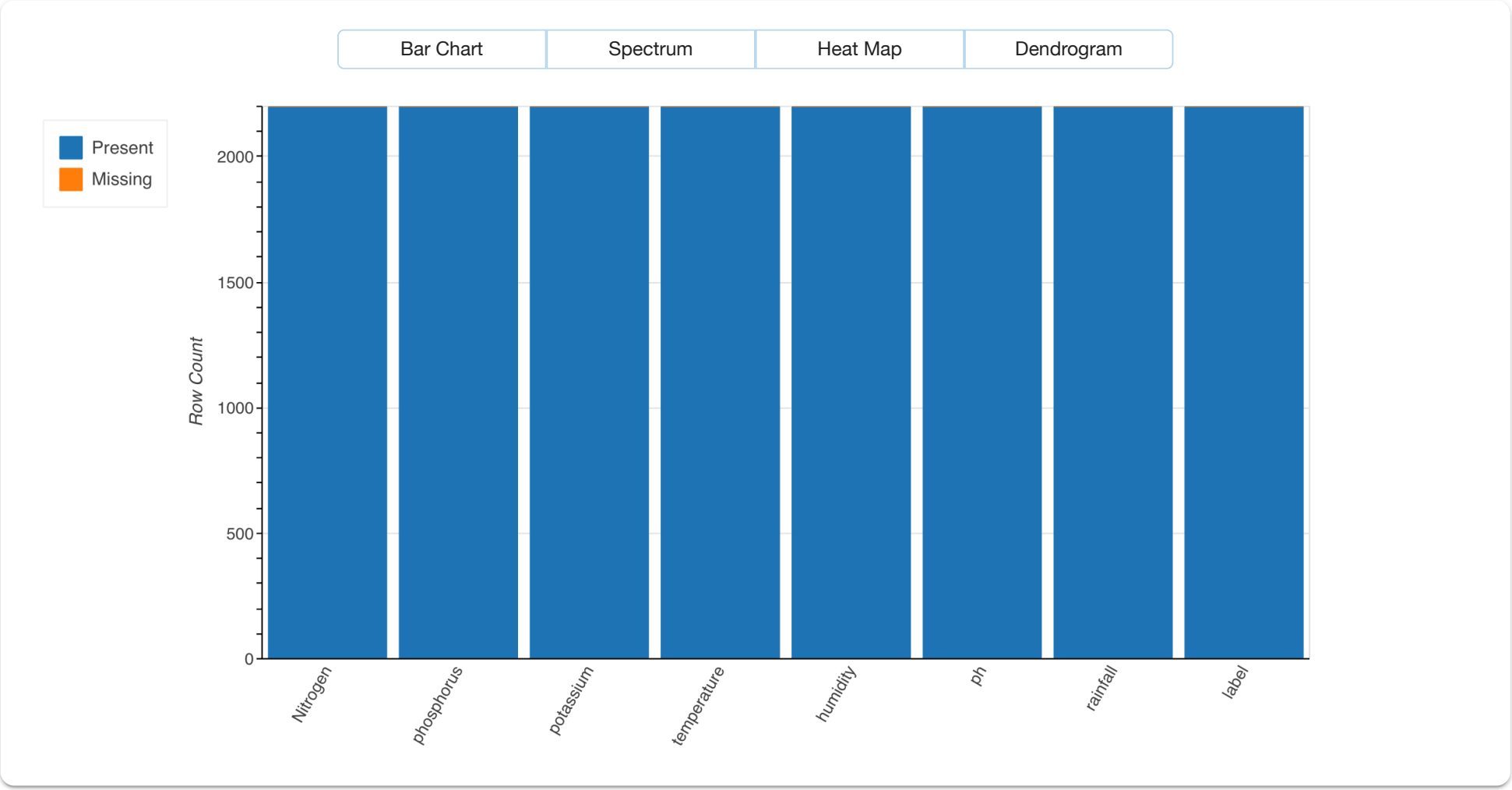




Correlations



Missing Values



Report generated with DataPrep

DATA SPLITTING

```
X = crop.drop('label',axis=1)
Y = crop['label']
```

```
X.head()
```

	Nitrogen	phosphorus	potassium	temperature	humidity	ph	rainfall
0	90	42	43	20.879744	82.002744	6.502985	202.935536
1	85	58	41	21.770462	80.319644	7.038096	226.655537
2	60	55	44	23.004459	82.320763	7.840207	263.964248
3	74	35	40	26.491096	80.158363	6.980401	242.864034
4	78	42	42	20.130175	81.604873	7.628473	262.717340

FEATURE SELECTION

```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
```

```
ordered_rank_features=SelectKBest(score_func=chi2,k=7)
ordered_feature=ordered_rank_features.fit(X,Y)
```

```
dtscores=pd.DataFrame(ordered_feature.scores_,columns=["Score"])
dtcolumns=pd.DataFrame(X.columns)
```

```
features_rank=pd.concat([dtcolumns,dtscores],axis=1)
```

```
features_rank.columns=['Features','Score']
features_rank
```

	Features	Score
0	Nitrogen	53144.698042
1	phosphorus	42500.133699
2	potassium	116710.530813
3	temperature	1092.422417
4	humidity	14755.489757
5	ph	74.886567
6	rainfall	54808.131541

```
features_rank.nlargest(10,'Score')
```

	Features	Score
2	potassium	116710.530813
6	rainfall	54808.131541
0	Nitrogen	53144.698042
1	phosphorus	42500.133699
4	humidity	14755.489757
3	temperature	1092.422417
5	ph	74.886567

Feature Importance

```
from sklearn.ensemble import ExtraTreesClassifier
#import matplotlib.pyplot as plt
model=ExtraTreesClassifier()
model.fit(X,Y)
```