dataset_exploration

December 13, 2025

1 Dataset Exploration

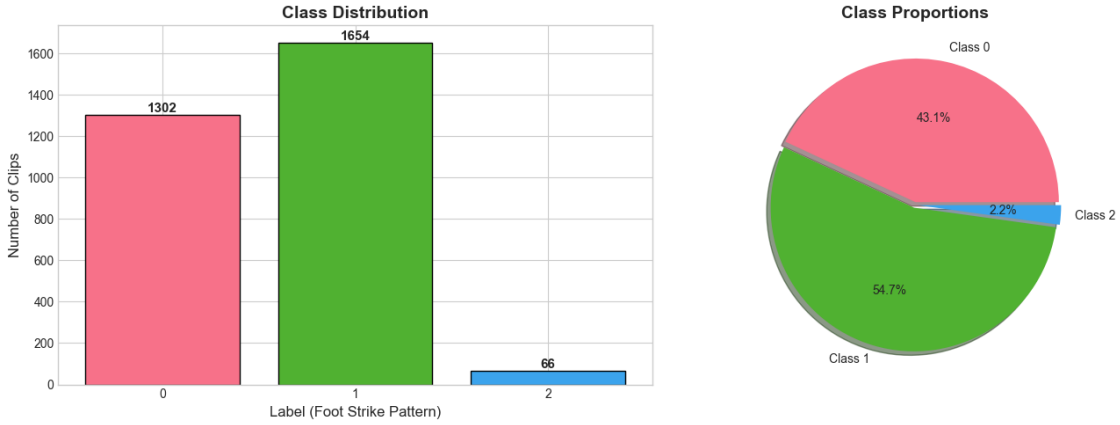
This notebook analyzes the running pattern dataset to understand: 1. **Class distribution** - Are labels balanced? 2. **Runner distribution** - Is any runner over-represented? 3. **Camera distribution** - Are camera angles balanced? 4. **Cross-correlations** - Is a specific class only from one camera/runner?

```
Total files: 3022
Parsed: 3022
Failed: 0

Unique runners: 17
Unique labels: [0, 1, 2]
Unique cameras: ['CAM1', 'CAM2', 'CAM3']
```

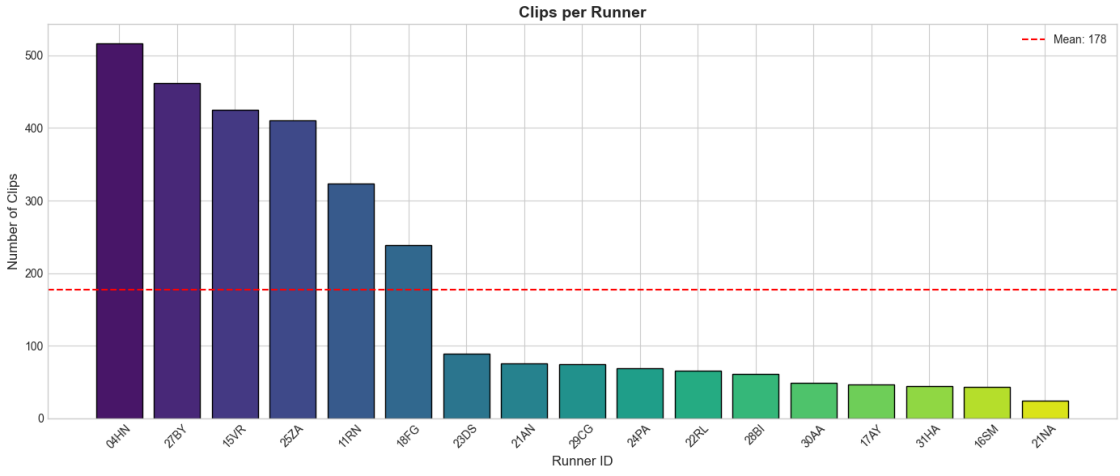
	runner	run	camera	lap	cut	label	path
0	04HN	2	CAM1	6	1	1	/Users/nirgofman/Desktop/running-pattern/final…
1	04HN	2	CAM1	6	2	1	/Users/nirgofman/Desktop/running-pattern/final…
2	04HN	2	CAM1	6	3	1	/Users/nirgofman/Desktop/running-pattern/final…
3	04HN	2	CAM1	6	4	1	/Users/nirgofman/Desktop/running-pattern/final…
4	04HN	2	CAM1	6	5	1	/Users/nirgofman/Desktop/running-pattern/final…
5	04HN	2	CAM1	6	6	1	/Users/nirgofman/Desktop/running-pattern/final…
6	04HN	2	CAM1	6	7	1	/Users/nirgofman/Desktop/running-pattern/final…
7	04HN	2	CAM1	7	1	1	/Users/nirgofman/Desktop/running-pattern/final…
8	04HN	2	CAM1	7	2	1	/Users/nirgofman/Desktop/running-pattern/final…
9	04HN	2	CAM1	7	3	1	/Users/nirgofman/Desktop/running-pattern/final…

1.1 1. Class Distribution (Label Imbalance)



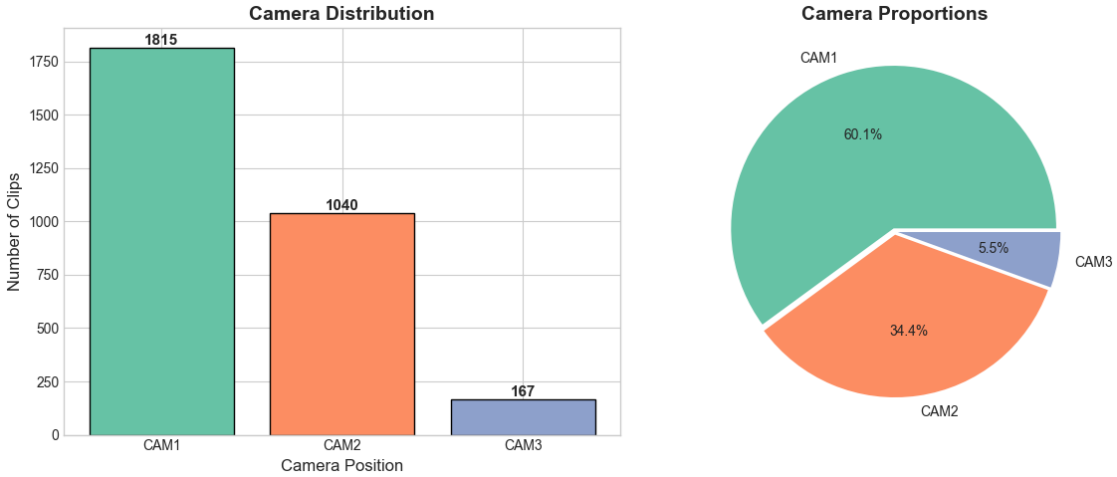

```
Class Imbalance Ratio: 25.06x
  WARNING: Severe class imbalance detected! Consider weighted loss or
oversampling.
```

1.2 2. Runner Distribution (Subject Imbalance)



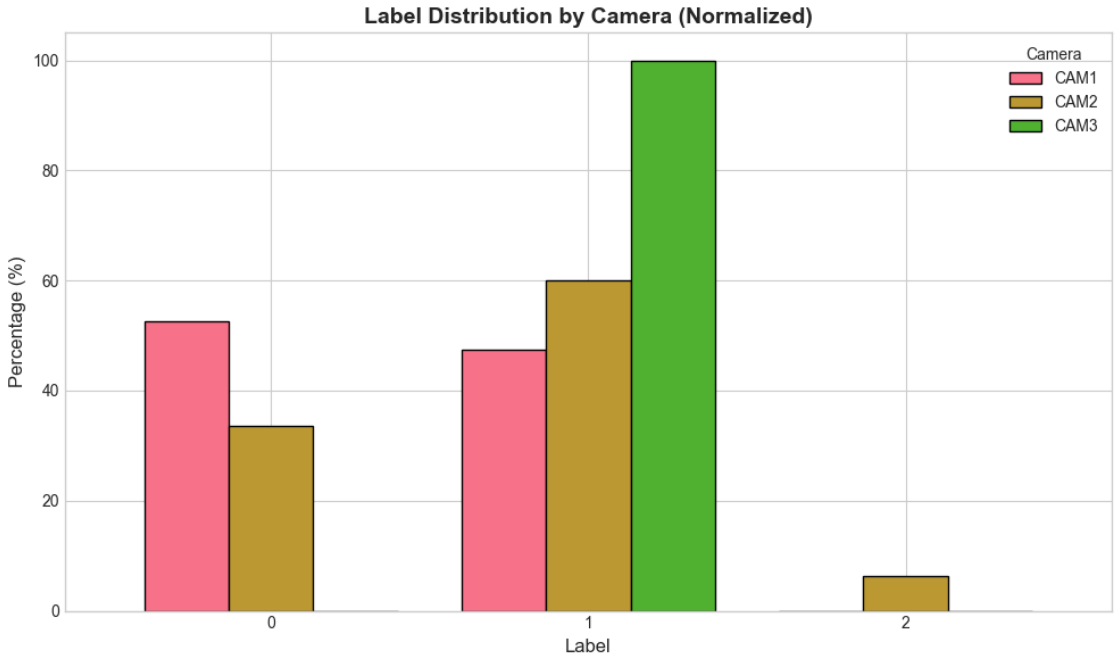

```
Runner count range: 24 - 517
Mean clips per runner: 177.8
Runner imbalance ratio: 21.54x
```

1.3 3. Camera Distribution

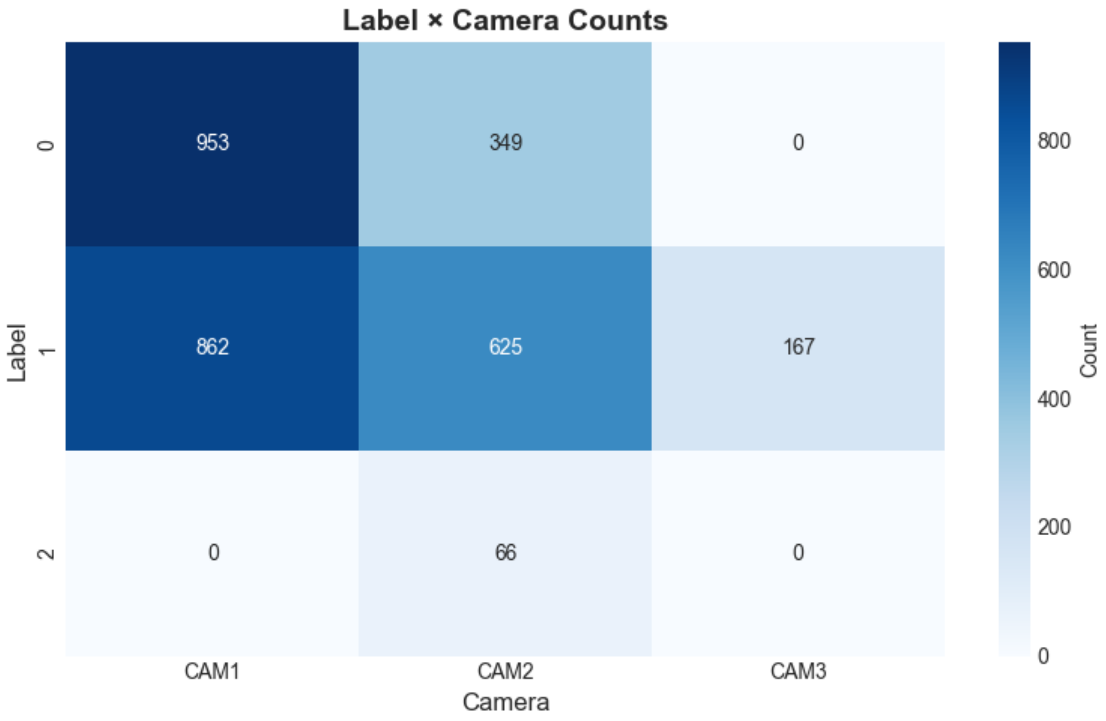


1.4 4. Cross-Correlations (Detecting Confounds)

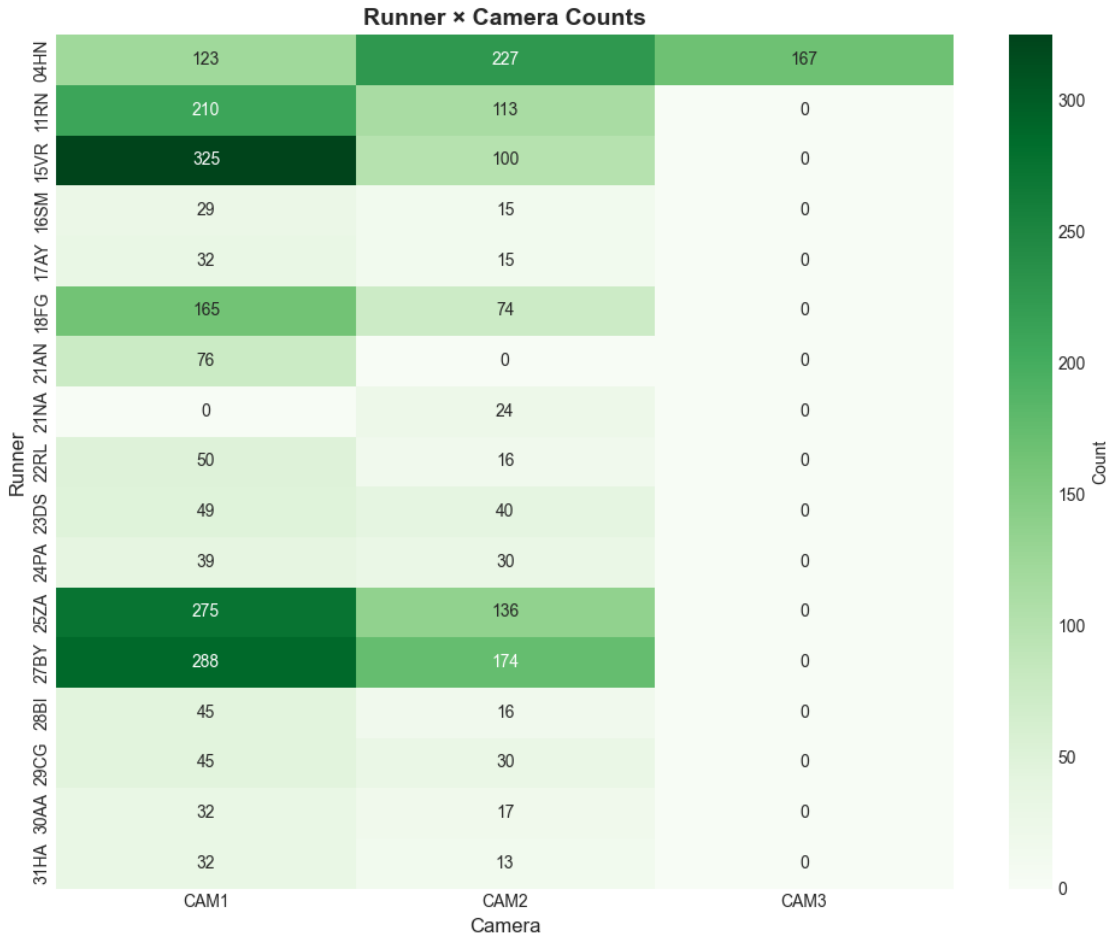
Check if labels are correlated with cameras or specific runners.



`If bars are similar across cameras, labels are NOT confounded with camera angle.`


Label × Camera Counts
Label
0
1
2
953
349
0
862
625
167
0
66
0
CAM1
CAM2
CAM3
Camera
Count
800
600
400
200
0


```
  5 runners have multiple labels:
runner
11RN    2
18FG    2
25ZA    2
27BY    2
30AA    2
dtype: int64
```



1.5 5. Summary Statistics

```
============================================================
DATASET SUMMARY
============================================================
Total clips: 3022
Unique runners: 17
Unique labels: 3
Cameras: ['CAM1', 'CAM2', 'CAM3']

Class Distribution:
  Label 0:  1302 clips ( 43.1%)
  Label 1:  1654 clips ( 54.7%)
  Label 2:    66 clips (  2.2%)

Class imbalance ratio: 25.06x
Runner imbalance ratio: 21.54x
```

```
============================================================
RECOMMENDATIONS
============================================================
  Severe class imbalance - use weighted loss or oversampling
   Large runner imbalance - ensure proper train/test split by runner

  Remember: Split by RUNNER ID, not randomly!
```