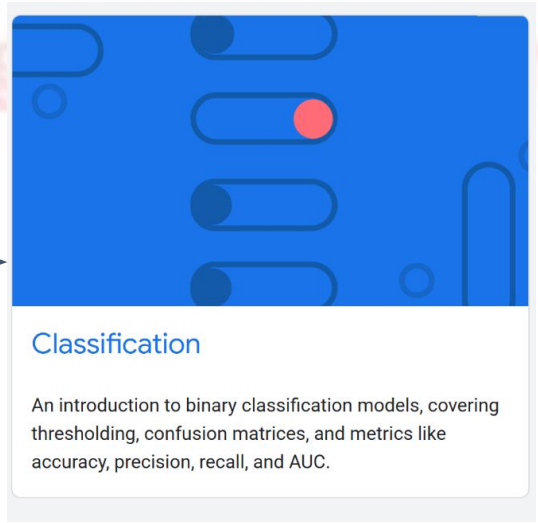
TECHCRUSH ARTIFICIAL INTELLIGENCE BOOTCAMP

Facilitator: Hammed Obasekore
September 10th, 2025

Recap



Understanding Data

Numerical Data

Means integers or floating-point values that behave like numbers

Examples;

- Temperature,
- Age,
- Height

Categorical Data

Means numbers that behave like categories

Examples;

- Postal Code,
- Dialing Code
  - (+233) + (+1) = +234
  - Ghana + USA = Nigeria ?

Understanding Data

Numerical Data

Raw Data

	Area	Perimete	Major_Axis_Lengt	Minor_Axis_Lengt	Eccentricity	Convex_Are	Extent	Class
2	15231	525.578979	229.7498779	85.09378815	0.928882003	15617	0.5728955	Cammeo
3	14656	494.311005	206.0200653	91.73097229	0.895404994	15072	0.6154363	Cammeo
4	14634	501.122009	214.106781	87.76828766	0.912118077	14954	0.6932588	Cammeo
5	13176	458.342987	193.3373871	87.44839478	0.891860902	13368	0.640669	Cammeo
6	14688	507.166992	211.7433777	89.31245422	0.906690896	15262	0.6460239	Cammeo
7	13479	477.015991	200.0530548	86.65029144	0.901328325	13786	0.6578973	Cammeo
8	15757	509.281006	207.2966766	98.33613586	0.88032347	16150	0.5897081	Cammeo
9	16405	526.570007	221.6125183	95.43670654	0.902520597	16837	0.6588883	Cammeo
10	14534	483.640991	196.6508179	95.05068207	0.875428557	14932	0.6496513	Cammeo
11	13485	471.570007	198.272644	87.72728729	0.896789312	13734	0.5723199	Cammeo
12	14930	499.924988	212.2458191	90.01747894	0.905606449	15248	0.6243727	Cammeo
13	14626	496.585999	204.5341339	92.97486877	0.890711546	15070	0.5702145	Cammeo
14	15926	522.73999	225.7360535	91.05709076	0.915033162	16240	0.7797689	Cammeo
15	14076	479.677002	199.489151	90.70998383	0.890638888	14434	0.7812188	Cammeo
16	13500	476.915009	202.5466766	85.4054718	0.906754851	13800	0.7177033	Cammeo
17	14349	496.946014	213.5440216	86.16077423	0.914988399	14678	0.6668371	Cammeo
18	15209	496.565002	214.0500793	91.02632141	0.90507257	15395	0.5693696	Cammeo
19	15238	496.871002	208.5317841	93.82839966	0.893054903	15487	0.7323145	Cammeo

Numerical Data

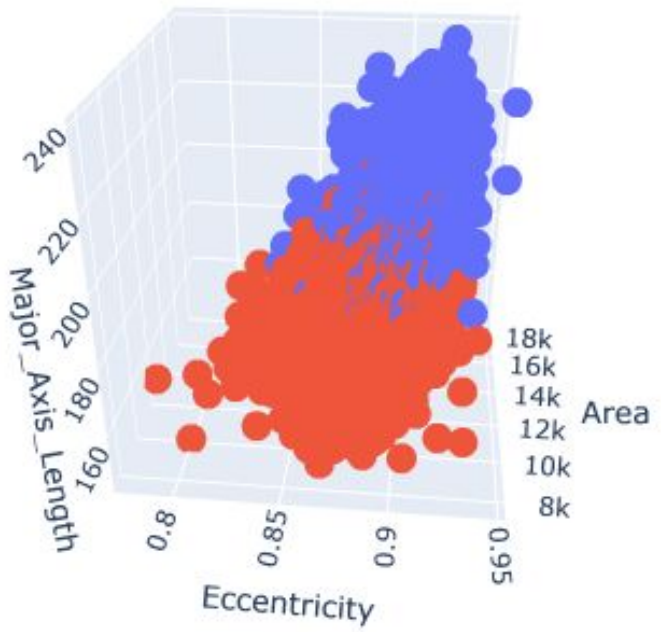
Feature Engineering, e.g

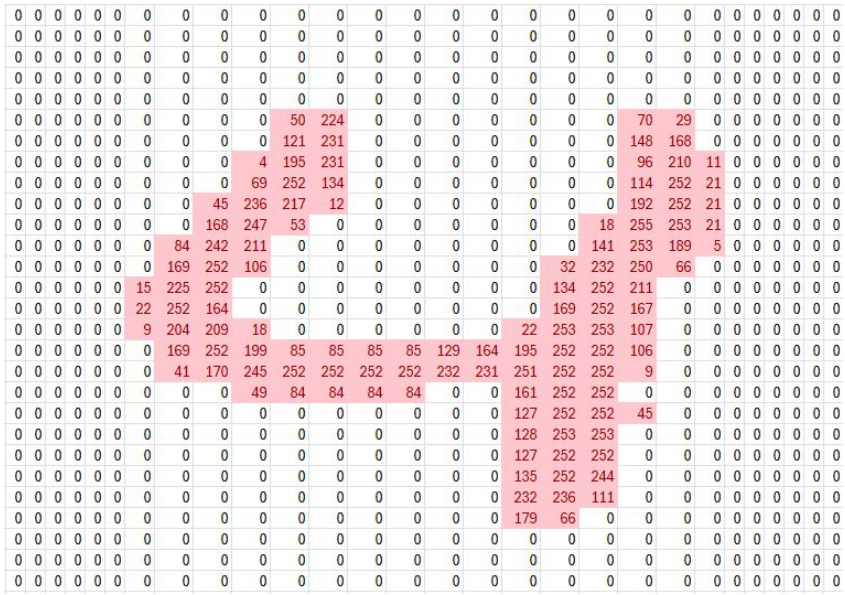
- Normalization: Converting numerical values into a standard range.
- Binning (also referred to as bucketing): Converting numerical values into buckets of ranges.

Numerical Data

Explore Data

- Visualize your data
  - Scatter Plot
  - Histogram
  - 3D Plot
- Be Creative





Numerical Data

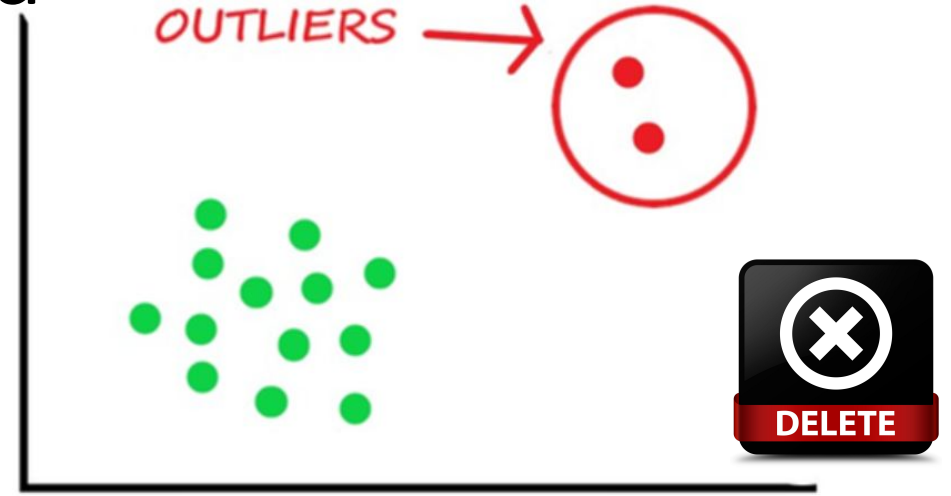
Explore Data

- Get statistics about your data
  - mean and median
  - standard deviation
  - percentiles

	Area	Perimeter	Major_Axis_Length	Minor_Axis_Length	Eccentricity	Convex_Area	Extent
count	3810.0	3810.0	3810.0	3810.0	3810.0	3810.0	3810.0
mean	12667.7	454.2	188.8	86.3	0.9	12952.5	0.7
std	1732.4	35.6	17.4	5.7	0.0	1777.0	0.1
min	7551.0	359.1	145.3	59.5	0.8	7723.0	0.5
25%	11370.5	426.1	174.4	82.7	0.9	11626.2	0.6
50%	12421.5	448.9	185.8	86.4	0.9	12706.5	0.6
75%	13950.0	483.7	203.6	90.1	0.9	14284.0	0.7
max	18913.0	548.4	239.0	107.5	0.9	19099.0	0.9

Numerical Data

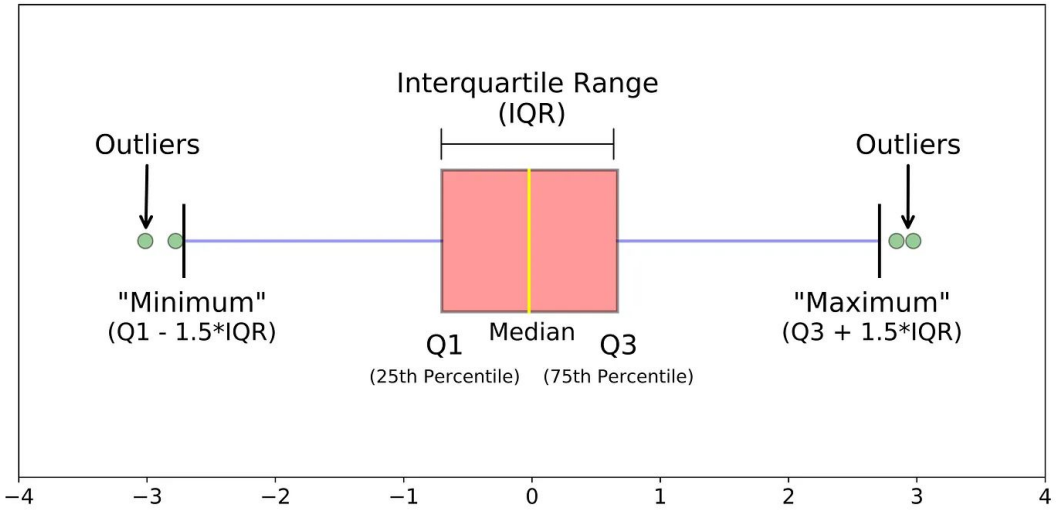
- Outliers a value distant from most other values in a feature or label.
  - cause problems in model training



outlier can be due to a mistake

- instrument malfunctioned
- experimenter mistake

Outlier may be a legitimate data point



Numerical Data

Normalization: to transform features to be on a similar scale

- linear scaling
- Z-score scaling
- log scaling

Use the following formula to scale to the standard range 0 to 1, inclusive:

$$x' = (x - x_{min})/(x_{max} - x_{min})$$

where:

- x' is the scaled value.
- x is the original value.
- x_{min} is the lowest value in the dataset of this feature.
- x_{max} is the highest value in the dataset of this feature.

Numerical Data

Normalization: to transform features to be on a similar scale

- linear scaling
- Z-score scaling
- log scaling

Use the following formula to normalize a value, x, to its Z-score:

$$x' = (x - \mu)/\sigma$$

where:

- x' is the Z-score.
- x is the raw value; that is, x is the value you are normalizing.
- μ is the mean.
- σ is the standard deviation.

Questions?