THE BUSINESS CLOUD™

EXTREME VALUE EXCLUSION: LOGIC & METHODS

DOMO AI Labs Teaching Assets

Introduction to Extreme Values

- Outliers
  - Observations that are atypically distant from other observations (statistically unusual)
- Unusual values / unreasonable values
  - Observations that are unusual / unreasonable for the given use case, regardless of their statistical properties

Employee ID	Hourly Wage
5	18
2	19
9	19
13	19
1	20
6	20
7	20
8	20
11	21
3	22
10	22
12	23
14	23
4	24
15	110

DOMO THE BUSINESS CLOUD™

Why Do We Care About Outliers?

- Because outliers….
  - Distort statistics about our data (e.g., means/averages, standard deviations, correlations, etc.)

Outlier Included?	Average Hourly Wage	Standard Deviation of Hourly Wage
Without Outlier	$20.71	$1.75
With Outlier	$26.67	$22.34

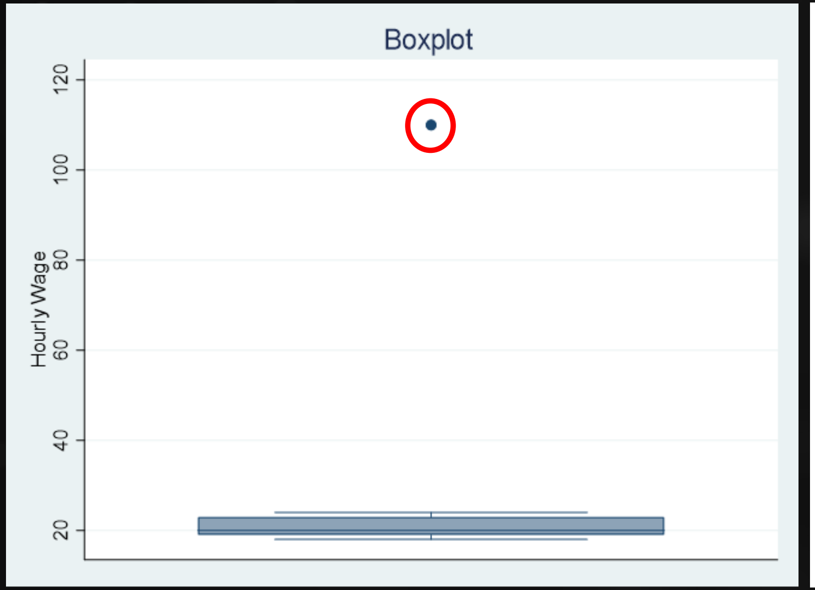
- As a result….
  - Hypothesis/significance tests (which are based on differences in means) between groups will be biased, which will subsequently affect model results
  - Larger standard deviation values reduces statistical power (which means hypothesis/significance tests will be less likely to detect an effect that exists), which will further affect model results

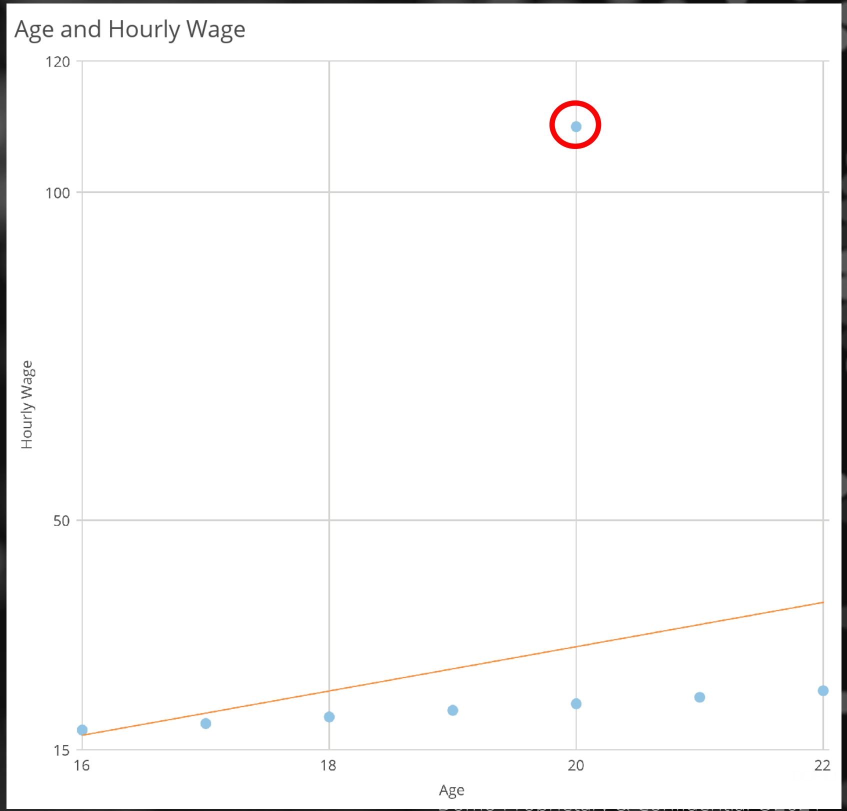
Methods for Outlier Detection

- Graphs & Visualizations
- Statistical methods:
  - Mean ± 3 Standard Deviations
    - Normalized/Scaled Transformation
  - Median Absolute Dispersion
  - Quartile-based Fences
  - Isolation Forest
- Business-Driven Caps/User-Defined Thresholds

Graphs & Visualizations

- **Graphs can *visually* highlight extreme values:** Boxplots, histograms, and scatterplots are commonly used for outlier detection
- **Pros of method:** Relatively easy to create and interpret graphs
- **Cons of method:** Outlier identification is subjective when using histograms and scatterplots (i.e., not based on statistical formula or criteria)





Mean ± 3 Standard Deviations

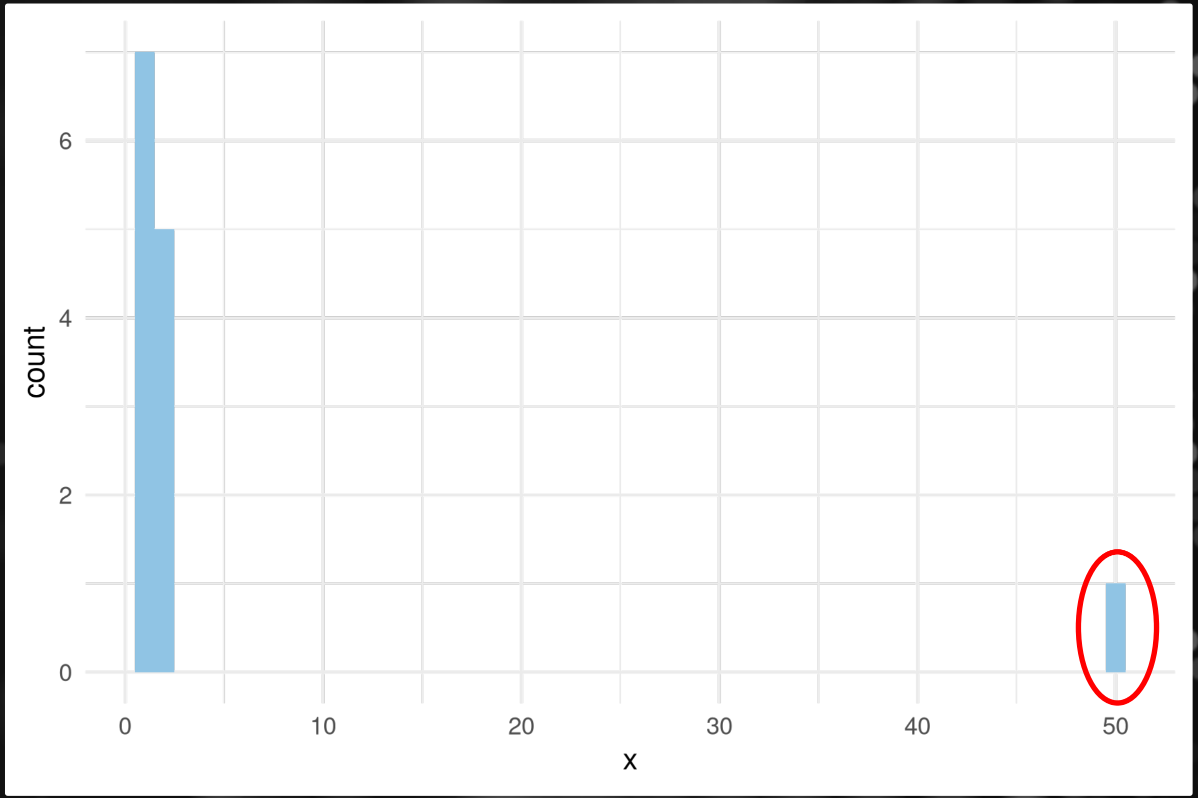
- Step 1: Calculate the mean & standard deviation
- Step 2: Calculate the upper and lower thresholds as:
  - Upper threshold: mean + (3 * standard deviation)
  - Lower threshold: mean - (3 * standard deviation)
- Step 3: Identify data points that are above or below the thresholds as outliers

- Optional: Standardize the data series with Z-scores, which will result in `(Mean=0, SD=1)`. This simplifies the outlier identification as any data point greater than 3 or less than –3 is an outlier

THE BUSINESS CLOUD™

Mean ± 3 Standard Deviations

- Data: [1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 50]
- Step 1: Mean=5.15, SD=13.48
- Step 2:
  - Lower = 5.15 - (3 * 13.48) = -35.29
  - Upper = 5.15 + (3 * 13.48) = 45.59
- Step 3: "50" is an outlier because 50 > 45.49



THE BUSINESS CLOUD™

Mean ± 3 Standard Deviations

- When to use this statistical method:
  - If your data is roughly normally distributed AND you have a large sample
- Pros of this statistical method:
  - It's relatively simple to implement
  - It's well-known and frequently used by statisticians
- Cons of this statistical method:
  - The mean and standard deviation values used in the calculation are strongly impacted by outliers. Therefore, this method is *fundamentally flawed*: it is supposed to guide our outlier detection, but the method itself is altered by the presence of extreme values

DOMO THE BUSINESS CLOUD™

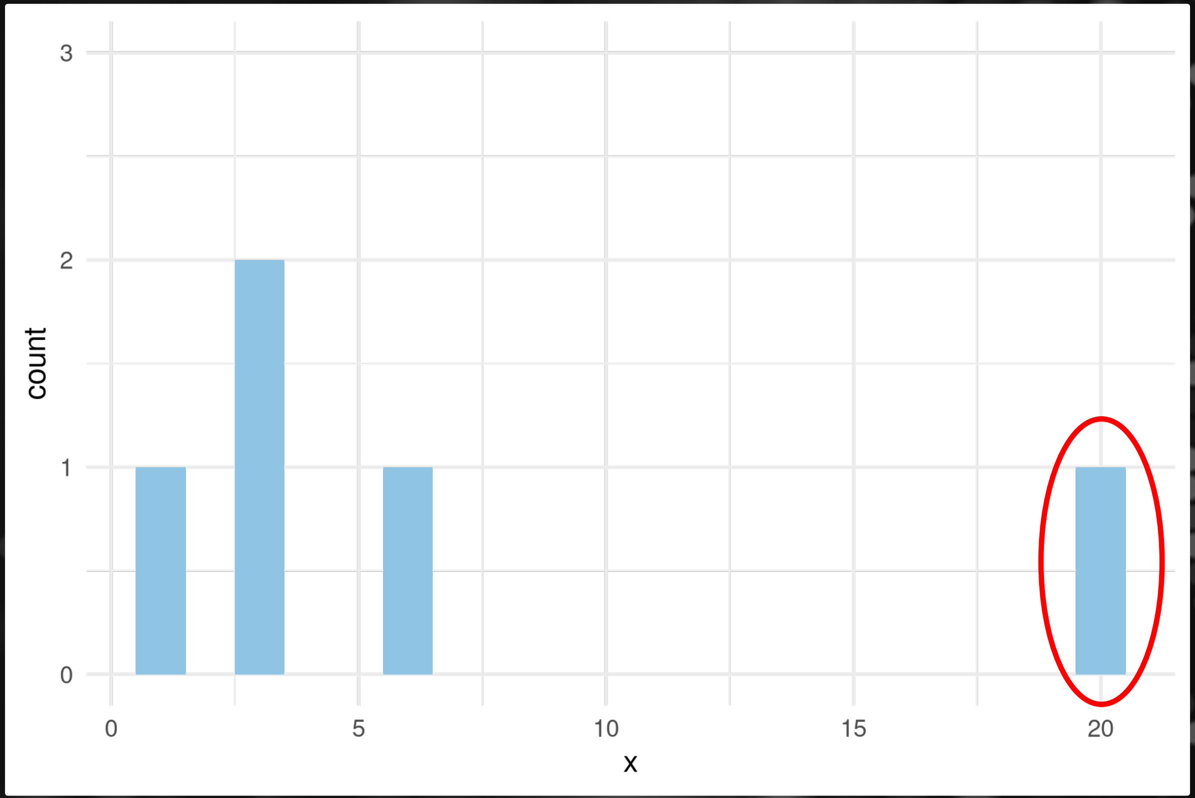
Median Absolute Dispersion

- Step 1: Calculate the median
- Step 2: Subtract the median from all data points
- Step 3: Take the absolute value of the data points from Step 2
- Step 4: Calculate the median of the Step 3 result
- Step 5: Multiply the median value obtained in Step 4 by 1.4826. This is a constant linked to the assumption of normality of the data, disregarding the abnormality induced by outliers
- Step 6: Calculate threshold values by taking the original median +/- (2.5 * Step 5 result)
- Step 7: Identify data points that are above or below the thresholds as outliers

THE BUSINESS CLOUD™

Median Absolute Dispersion

- Data: [1, 3, 3, 6, 20]
- Step 1: Median = 3
- Step 2: [-2, 0, 0, 3, 17]
- Step 3: [2, 0, 0, 3, 17]
- Step 4: Median = 2
- Step 5: M = 2 * 1.4826 = 2.9652
- Step 6:
  - Lower = 3 - (2.5 * 2.9652) = -4.413
  - Upper = 3 + (2.5 * 2.9652) = 10.413
- Step 7: "20" is an outlier because 20 > 10.413



DOMO THE BUSINESS CLOUD™

Median Absolute Dispersion

- Pros of this statistical method:
  - It is *not* reliant on calculating the mean and standard deviation values, which are both impacted by outliers
  - It can be implemented with a small or large sample
  - It's relatively simple to implement

THE BUSINESS CLOUD™

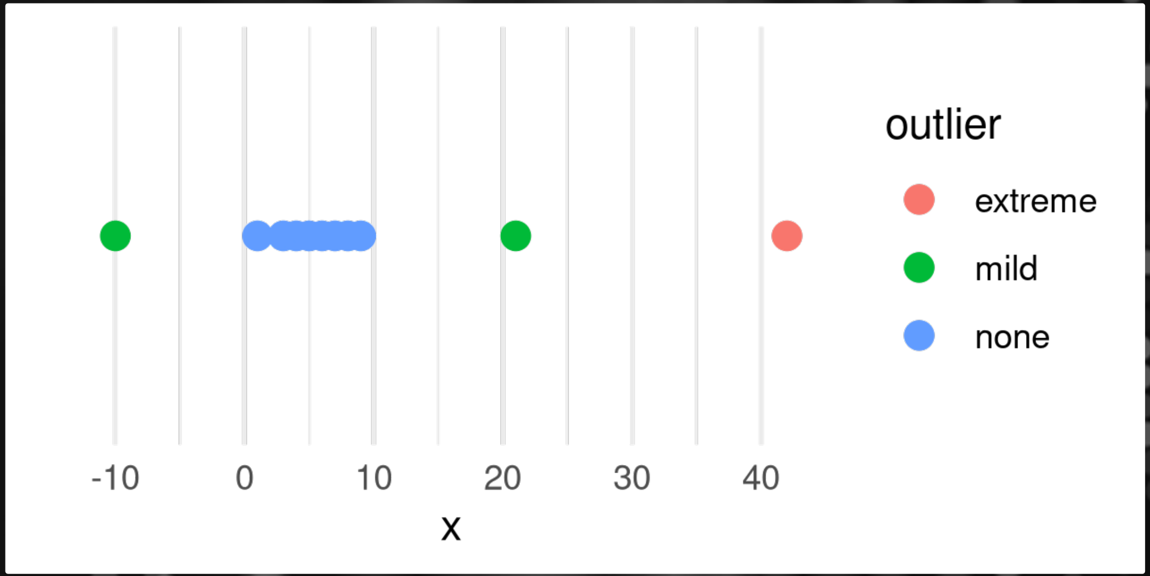
Quartile-based Fences

- Step 1: Calculate 1st and 3rd quartiles (25th and 75th percentiles)
- Step 2: Calculate the Interquartile Range as follows: `IQR = Q3 - Q1`
- Step 3: Calculate the following thresholds or "fences":
  - Lower inner fence `= Q1 - (IQR * 1.5)`
  - Lower outer fence `= Q1 - (IQR * 3)`
  - Upper inner fence `= Q3 + (IQR * 1.5)`
  - Upper outer fence `= Q3 + (IQR * 3)`
- Step 4: Classify data points beyond the inner fence as *mild* outliers, and data points beyond the outer fence as *extreme outliers*

DOMO THE BUSINESS CLOUD™

Quartile-based Fences

- Data: `[-10, 1, 3, 4, 5, 6, 7, 8, 9, 21, 42]`
- Step 1: `Q1 = 3.5`, `Q3 = 8.5`
- Step 2: `IQR = 8.5 - 3.5 = 5`
- Step 3:
  - Lower inner fence `= 3.5 - (5 * 1.5) = -4`
  - Lower outer fence `= 3.5 - (5 * 3) = -11.5`
  - Upper inner fence `= 8.5 + (5 * 1.5) = 16`
  - Upper outer fence `= 8.5 + (5 * 3) = 23.5`
- Step 4:
  - Classify -10 and 21 as *mild outliers*
  - Classify 42 as an *extreme outlier*



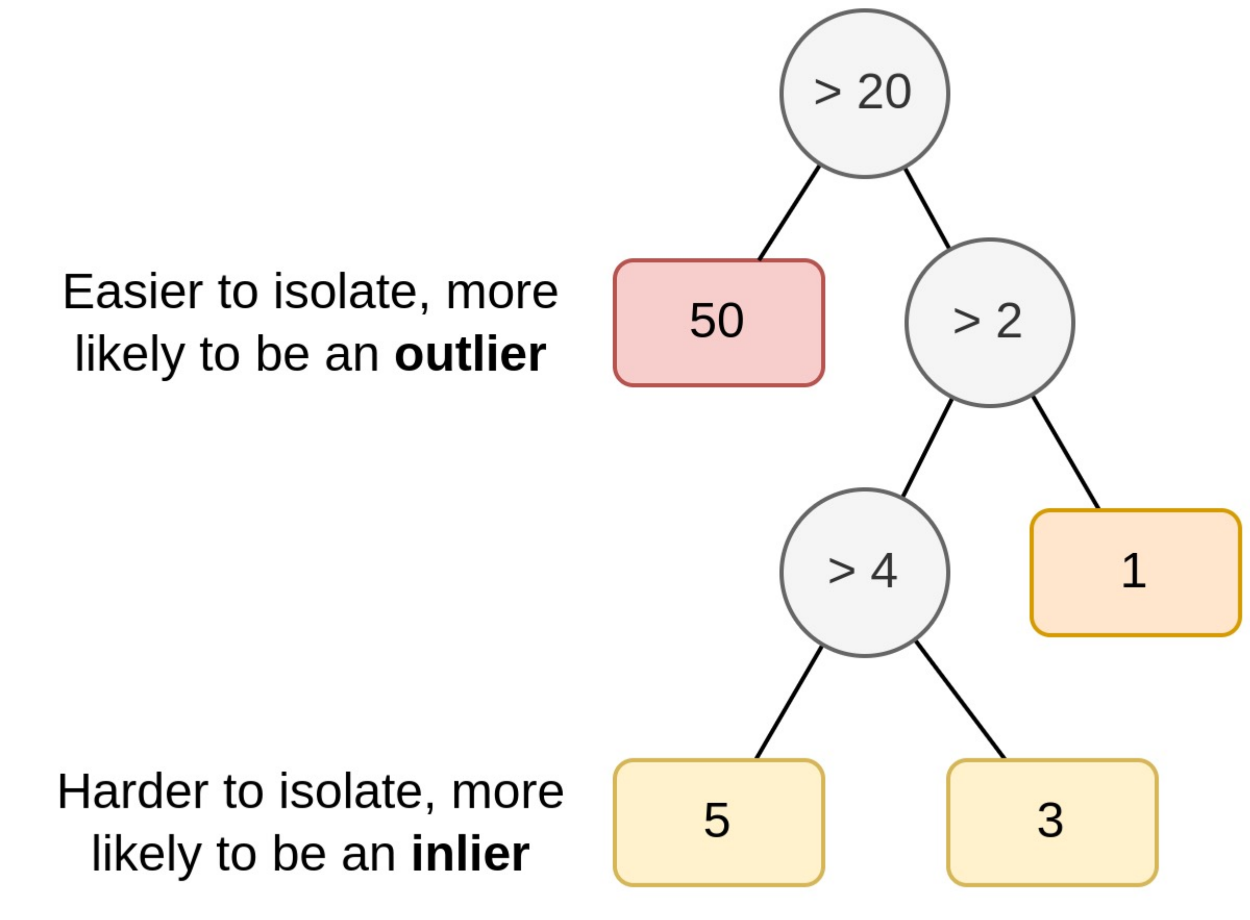
Quartile-based Fences

- Pros of this statistical method:
  - It is *not* reliant on calculating the mean and standard deviation values, which are both impacted by outliers
  - It's relatively simple to implement

THE BUSINESS CLOUD™

Isolation Forest

- Algorithmic Approach:
  - Isolation Forests use binary-search trees to determine the "anomaly score" of each data point
  - Data points that are easier to isolate in the tree are more likely to be outliers



Isolation Forest

- Code Sample:
  - The Isolation Forest algorithm is available in Python and R via open-source libraries

```
[1]: import numpy as np
     from sklearn.ensemble import IsolationForest

     x = np.array([1, 1, 3, 5, 50])
     x = x.reshape(-1, 1)

     # -1 indicates predicted outliers
     clf = IsolationForest().fit(x)
     clf.predict(x)

[1]: array([ 1,  1,  1,  1, -1])
```

THE BUSINESS CLOUD™

Isolation Forest

- Pros:
  - Can identify multi-dimensional outliers
- Cons:
  - Requires programming experience / environment
  - Results may vary with algorithm parameters
  - More of a "black-box" than other methods

DOMO THE BUSINESS CLOUD™

Business-Driven Caps / User-Defined Thresholds

- **Business-Driven Caps:**
  - Limit data to appropriate ranges required by the business use-case
  - Example: If you are modeling car prices, you may consider removing cars from your analysis that are above a certain price. These high-end cars might not be true statistical outliers, but if your business-case is focused on modeling everyday automobiles, selecting an upper limit for price will be beneficial
- **Common Sense:**
  - Other thresholds may simply be the result of common-sense or "gut feelings"
  - This approach is often included in Data Quality Assurance Analysis
  - Example: If looking at the weights of mice, any negative value should be treated as an extreme value, regardless of whether these values were identified as outliers with statistical methods

Best Methods for Identifying Statistical Outliers

- Understand the size of your data and the distribution of your population
  - Methods like Mean+3SD work best with a large sample and an underlying normal distribution
- Use a combination of graphical and statistical methods to validate results
- Compare the results of identification methods
  - Data points flagged via multiple methods are likely to be statistical outliers

DOMO THE BUSINESS CLOUD™

Best Methods for Identifying Practical Outliers

- Understand the business problem
  - Whether or not extreme values are removed from analysis will depend on the use-case
- Iterate with stakeholders and to evaluate the appropriate thresholds
  - Different stakeholders (especially the data experts) can provide background information on extreme values, and help identify appropriate thresholds

THE BUSINESS CLOUD™

Next Steps

- How do you handle outliers after identifying them?
  - Drop the observation from your sample
    - Be careful, as this approach reduces your sample size
  - Substitute the value of the outlier data point with a value that is less extreme
    - e.g., substitute outliers with the mean, 90^{th} percentile value, etc
  - Do nothing
    - If you only have a few outliers relative to your sample size, they may not have a large influence on your results
    - Evaluate your summary statistics and modeling results with and without outliers to confirm their influence
- Your approach will vary from one analysis to the next

THE BUSINESS CLOUD™

QUESTIONS?