# On the Outlier Detection for Standardized Tests

Nicolás Acevedo[1], Charles Thraves[1], and María Leonor Varas[2]

Universidad de Chile
[1]: Departamento de Ingeniería Industrial
[2]: Departamento de Ingeniería Matemática and DEMRE

March 7, 2022

# Motivation

## PDT (former PSU)

- Chilean standardized college admission test (similar to SAT in the US)

# Motivation

## PDT (former PSU)

- Chilean standardized college admission test (similar to SAT in the US)
- 4 multiple choice tests:
  1. Language (LANG)
  2. Math (MATH)
  3. History (HIST) or Sciences (SCI)

## PDT (former PSU)

- Chilean standardized college admission test (similar to SAT in the US)
- 4 multiple choice tests:
  1. Language (LANG)
  2. Math (MATH)
  3. History (HIST) or Sciences (SCI)
- $\geq 230,000$ applicants per year (Chile's population is almost 20 million)

# Motivation

## PDT (former PSU)

- Chilean standardized college admission test (similar to SAT in the US)
- 4 multiple choice tests:
  1. Language (LANG)
  2. Math (MATH)
  3. History (HIST) or Sciences (SCI)
- $\geq 230,000$ applicants per year (Chile's population is almost 20 million)
- Some of the obtained scores may not be truthful (or correct)
  E.g.:
  - Correction was performed with the wrong answer key
  - Operational issue (pencil did not mark well)
  - Cheating
  - Other

# Motivation

- **Problem**: We do not know which scores are mistaken (as in supervised learning).

# Motivation

- **Problem**: We do not know which scores are mistaken (as in supervised learning).
- **Solution Approach**: Apply unsupervised machine learning techniques to find cases (i.e. test scores) worth to investigate.

# Motivation

- **Problem**: We do not know which scores are mistaken (as in supervised learning).
- **Solution Approach**: Apply unsupervised machine learning techniques to find cases (i.e. test scores) worth to investigate.
  - Not every combination of scores is equally likely to occur.
    E.g. (scores between 0 and 100):
    1. LANG/MATH/HIST $= 84/80/90$ 👍
    2. LANG/MATH/HIST $= 10/80/90$ 🤔

# Motivation

- **Problem**: We do not know which scores are mistaken (as in supervised learning).
- **Solution Approach**: Apply unsupervised machine learning techniques to find cases (i.e. test scores) worth to investigate.
  - Not every combination of scores is equally likely to occur. E.g. (scores between 0 and 100):
    1. LANG/MATH/HIST $= 84/80/90$ 👍
    2. LANG/MATH/HIST $= 10/80/90$ 🤔
  - We are looking for the less likely cases. It is an outlier detection problem.

# Motivation

- **Problem**: We do not know which scores are mistaken (as in supervised learning).
- **Solution Approach**: Apply unsupervised machine learning techniques to find cases (i.e. test scores) worth to investigate.
  - Not every combination of scores is equally likely to occur.
    E.g. (scores between 0 and 100):
    1. LANG/MATH/HIST $= 84/80/90$
    2. LANG/MATH/HIST $= 10/80/90$
  - We are looking for the less likely cases. It is an outlier detection problem.

**Research Question:**

**How can we identify outliers for standardized tests while providing a way to interpret and classify the type of outlier?**

# Summary

1 PDT Scores → 2 Outlier Detection Methods → 3 PCA → 4 $k$-means

# Summary

① **PDT Scores** → ② Outlier Detection Methods → ③ PCA → ④ $k$-means

# PDT Scores

For this presentation we consider:

- Applicants who took the tests LANG/MATH/HIST in December 2021.
  - $\geq$103,000 applicants on this group.

# PDT Scores

For this presentation we consider:

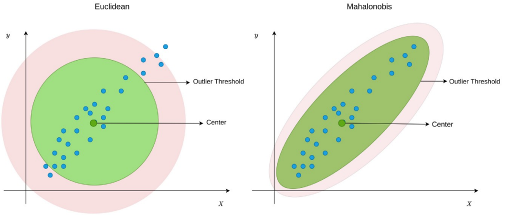
- Applicants who took the tests LANG/MATH/HIST in December 2021.
  - $\geq$103,000 applicants on this group.
- Test scores are the percentages of correct answers for each test.
  - Scores between $0$ and $100$.

# Summary

(1) PDT Scores $\rightarrow$ **(2) Outlier Detection Methods** $\rightarrow$ (3) PCA $\rightarrow$ (4) $k$-means

## Multivariate Outlier Detection Methods:

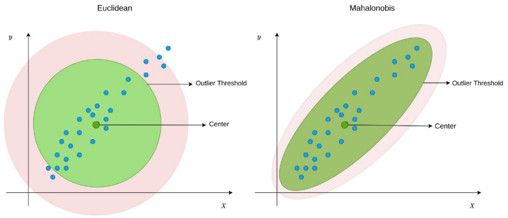
1. Mahalanobis Distance



# Outlier Detection Methods - Methodology

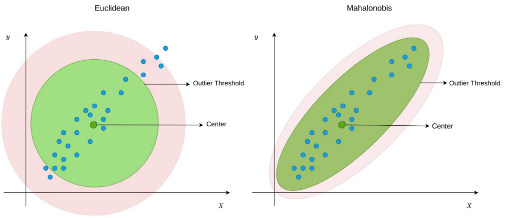
## Multivariate Outlier Detection Methods:

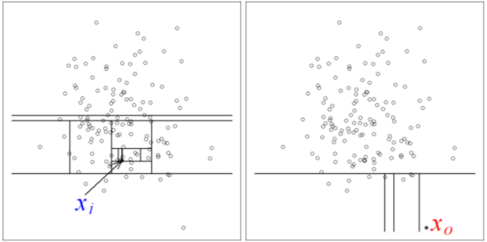
1. Mahalanobis Distance
2. Robust Mahalanobis Distance



## Multivariate Outlier Detection Methods:

1. Mahalanobis Distance
2. Robust Mahalanobis Distance
3. Isolation Forest (iForest)

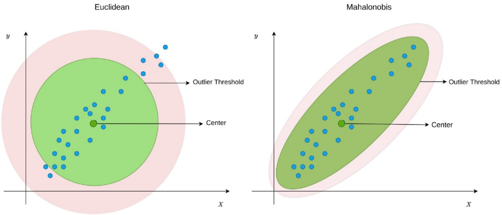


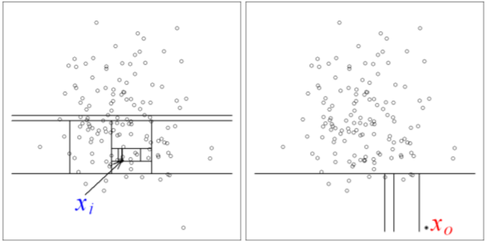




(a) Isolating $x_i$ (b) Isolating $x_o$

## Multivariate Outlier Detection Methods:

1. Mahalanobis Distance
2. Robust Mahalanobis Distance
3. Isolation Forest (iForest)
4. Extended iForest

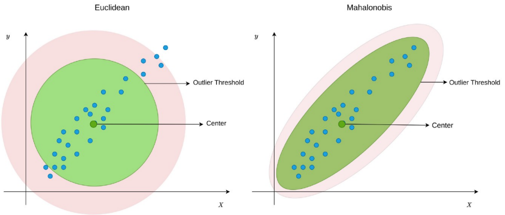


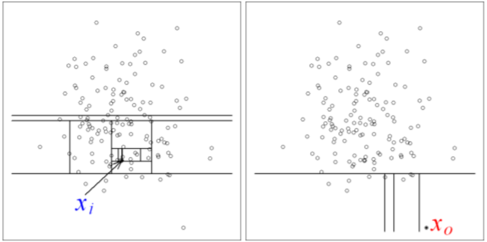




(a) Isolating $x_i$ (b) Isolating $x_o$

## Multivariate Outlier Detection Methods:

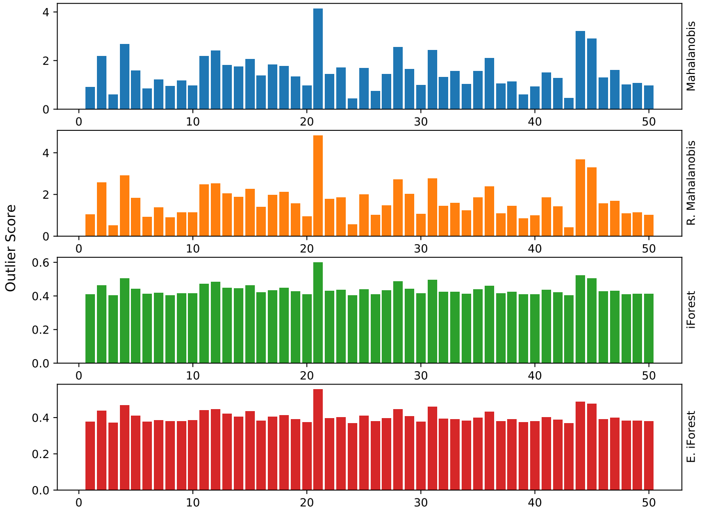
1. Mahalanobis Distance
2. Robust Mahalanobis Distance
3. Isolation Forest (iForest)
4. Extended iForest



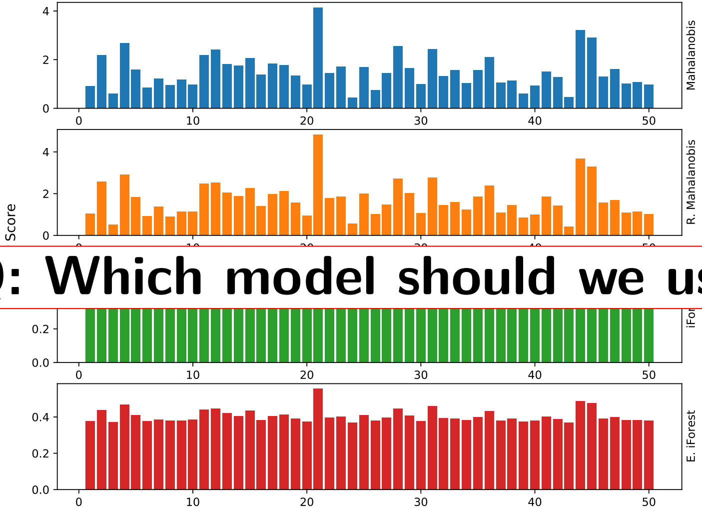




(a) Isolating $x_i$ (b) Isolating $x_o$

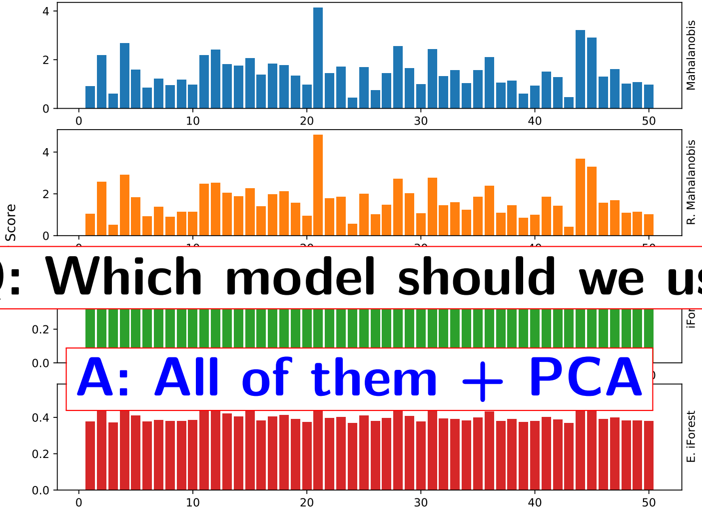
Every method gives us an outlier score (i.e. anomaly score) for all the applicants.


Outlier Score
Mahalanobis
R. Mahalanobis
iForest
E. iForest
0
2
4
0.0
0.2
0.4
0.6
0
10
20
30
40
50

# Outlier Detection Methods - Scorings Results




**Q: Which model should we use?**




**Q: Which model should we use?**

**A: All of them + PCA**

# Summary

① PDT Scores → ② Outlier Detection Methods → **③ PCA** → ④ $k$-means

# PCA - Results

Applying PCA to the scores given by the 4 methods:

| | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| Value | 3.76 | 0.19 | 0.04 | 0.01 |
| Percentage | 94.07% | 4.72% | 1.00% | 0.21% |

Explained variance by each principal component

# PCA - Results

Applying PCA to the scores given by the 4 methods:

| | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| Value | 3.76 | 0.19 | 0.04 | 0.01 |
| Percentage | 94.07% | 4.72% | 1.00% | 0.21% |

Explained variance by each principal component

| | Mahal. | R. Mahal. | iForest | E. iForest | PC1 |
|---|---|---|---|---|---|
| Mahalanobis | 1.00 | | | | |
| R. Mahalanobis | 0.80 | 1.00 | | | |
| iForest | 0.70 | 0.74 | 1.00 | | |
| E. iForest | 0.71 | 0.76 | 0.91 | 1.00 | |
| PC1 | 0.83 | 0.86 | 0.85 | 0.87 | 1.00 |

Correlation Matrix with Kendall's $\tau$ coefficient

# PCA - Results

| **Rank** | LANG | MATH | HIST | SCI | **PC1** |
|---|---|---|---|---|---|
| **1** | 13 | 20 | 98 | | **17.76** |
| **2** | 7 | 20 | 72 | | **14.90** |
| **3** | 28 | 22 | 95 | | **14.63** |
| **4** | 67 | 2 | 93 | | **13.49** |
| **5** | 38 | 27 | 97 | | **13.32** |
| **6** | 88 | 100 | 48 | | **13.23** |
| **7** | 55 | 97 | 50 | 28 | **12.87** |
| **8** | 67 | 82 | 18 | | **12.83** |
| **9** | 58 | 97 | 45 | | **12.79** |
| **10** | 28 | 68 | 67 | | **12.66** |
| **11** | 55 | 97 | 67 | 61 | **12.64** |
| **12** | 38 | 80 | 70 | 26 | **12.54** |
| **13** | 18 | 58 | 7 | | **12.52** |
| **14** | 70 | 90 | 30 | | **12.49** |
| **15** | 37 | 82 | 60 | 60 | **12.47** |
| **16** | 62 | 73 | 13 | | **12.35** |
| **17** | 68 | 12 | 98 | | **12.21** |
| **18** | 63 | 93 | 38 | | **12.14** |
| **19** | 55 | 83 | 92 | | **12.13** |
| **20** | 98 | 100 | 100 | | **12.12** |

# PCA - Results

| Rank | LANG | MATH | HIST | SCI | PC1 |
|---|---|---|---|---|---|
| **1** | **13** | **20** | **98** | | **17.76** |
| **2** | **7** | **20** | **72** | | **14.90** |
| **3** | **28** | **22** | **95** | | **14.63** |
| **4** | 67 | 2 | 93 | | **13.49** |
| **5** | **38** | **27** | **97** | | **13.32** |
| **6** | 88 | 100 | 48 | | **13.23** |
| **7** | 55 | 97 | 50 | 28 | **12.87** |
| **8** | 67 | 82 | 18 | | **12.83** |
| **9** | 58 | 97 | 45 | | **12.79** |
| **10** | 28 | 68 | 67 | | **12.66** |
| **11** | 55 | 97 | 67 | 61 | **12.64** |
| **12** | 38 | 80 | 70 | 26 | **12.54** |
| **13** | 18 | 58 | 7 | | **12.52** |
| **14** | 70 | 90 | 30 | | **12.49** |
| **15** | 37 | 82 | 60 | 60 | **12.47** |
| **16** | 62 | 73 | 13 | | **12.35** |
| **17** | 68 | 12 | 98 | | **12.21** |
| **18** | 63 | 93 | 38 | | **12.14** |
| **19** | 55 | 83 | 92 | | **12.13** |
| **20** | 98 | 100 | 100 | | **12.12** |

# PCA - Results

| **Rank** | LANG | MATH | HIST | SCI | **PC1** |
|---|---|---|---|---|---|
| **1** | **13** | **20** | **98** | | **17.76** |
| **2** | **7** | **20** | **72** | | **14.90** |
| **3** | **28** | **22** | **95** | | **14.63** |
| **4** | 67 | 2 | 93 | | **13.49** |
| **5** | **38** | **27** | **97** | | **13.32** |
| **6** | **88** | **100** | **48** | | **13.23** |
| **7** | 55 | 97 | 50 | 28 | **12.87** |
| **8** | **67** | **82** | **18** | | **12.83** |
| **9** | 58 | 97 | 45 | | **12.79** |
| **10** | 28 | 68 | 67 | | **12.66** |
| **11** | 55 | 97 | 67 | 61 | **12.64** |
| **12** | 38 | 80 | 70 | 26 | **12.54** |
| **13** | 18 | 58 | 7 | | **12.52** |
| **14** | **70** | **90** | **30** | | **12.49** |
| **15** | 37 | 82 | 60 | 60 | **12.47** |
| **16** | **62** | **73** | **13** | | **12.35** |
| **17** | 68 | 12 | 98 | | **12.21** |
| **18** | **63** | **93** | **38** | | **12.14** |
| **19** | 55 | 83 | 92 | | **12.13** |
| **20** | 98 | 100 | 100 | | **12.12** |

# PCA - Results

| **Rank** | LANG | MATH | HIST | SCI | **PC1** |
|---:|---:|---:|---:|---:|---:|
| **1** | **13** | **20** | **98** | | **17.76** |
| **2** | **7** | **20** | **72** | | **14.90** |
| **3** | **28** | **22** | **95** | | **14.63** |
| **4** | 67 | 2 | 93 | | **13.49** |
| **5** | **38** | **27** | **97** | | **13.32** |
| **6** | **88** | **100** | **48** | | **13.23** |
| **7** | 55 | 97 | 50 | 28 | **12.87** |
| **8** | **67** | **82** | **18** | | **12.83** |
| **9** | 58 | 97 | 45 | | **12.79** |
| **10** | 28 | 68 | 67 | | **12.66** |
| **11** | 55 | 97 | 67 | 61 | **12.64** |
| **12** | 38 | 80 | 70 | 26 | **12.54** |
| **13** | 18 | 58 | 7 | | **12.52** |
| **14** | **70** | **90** | **30** | | **12.49** |
| **15** | 37 | 82 | 60 | 60 | **12.47** |
| **16** | **62** | **73** | **13** | | **12.35** |
| **17** | 68 | 12 | 98 | | **12.21** |
| **18** | **63** | **93** | **38** | | **12.14** |
| **19** | 55 | 83 | 92 | | **12.13** |
| **20** | **98** | **100** | **100** | | **12.12** |

# Summary

1 PDT Scores → 2 Outlier Detection Methods → 3 PCA → **4 *k*-means**

# Results - Clustering

In order to group the different patterns, we:

- Took 1% of students with highest scores
- Clustered according to their scores deviations from their applicants' average score

| **Cluster** | LANG | MATH | HIST |
|---|---|---|---|
| **1** | 6.96 | 15.08 | -22.04 |
| **2** | 17.62 | -42.01 | 24.39 |
| **3** | -13.09 | 4.07 | 9.02 |
| **4** | -7.78 | -23.80 | 31.58 |
| **5** | -5.91 | 16.53 | -10.63 |
| **6** | 26.03 | -4.23 | -21.79 |
| **7** | -2.27 | 2.82 | -0.55 |

*k*-means centroids

# Results - Clustering

| Cluster | LANG | MATH | HIST | PC1 |
|---|---|---|---|---|
| | 88 | 100 | 48 | **13.23** |
| | 67 | 82 | 18 | **12.83** |
| **1** | 70 | 90 | 30 | **12.49** |
| | 62 | 73 | 13 | **12.35** |
| | 63 | 93 | 38 | **12.14** |
| | 67 | 2 | 93 | **13.49** |
| | 68 | 12 | 98 | **12.21** |
| **2** | 93 | 22 | 97 | **10.09** |
| | 90 | 18 | 93 | **9.64** |
| | 78 | 18 | 95 | **9.51** |
| | 28 | 68 | 67 | **12.66** |
| | 38 | 80 | 70 | **12.54** |
| **3** | 55 | 83 | 92 | **12.13** |
| | 80 | 100 | 97 | **11.77** |
| | 2 | 2 | 28 | **11.29** |
| | 13 | 20 | 98 | **17.76** |
| | 7 | 20 | 72 | **14.90** |
| **4** | 28 | 22 | 95 | **14.63** |
| | 38 | 27 | 97 | **13.32** |
| | 23 | 12 | 67 | **11.25** |

| Cluster | LANG | MATH | HIST | PC1 |
|---|---|---|---|---|
| | 55 | 97 | 50 | **12.87** |
| | 58 | 97 | 45 | **12.79** |
| **5** | 55 | 97 | 67 | **12.64** |
| | 18 | 58 | 7 | **12.52** |
| | 37 | 82 | 60 | **12.47** |
| | 78 | 65 | 23 | **9.74** |
| | 77 | 48 | 17 | **8.88** |
| **6** | 93 | 60 | 40 | **8.84** |
| | 92 | 50 | 33 | **8.63** |
| | 78 | 23 | 13 | **8.41** |
| | 98 | 100 | 100 | **12.12** |
| | 93 | 100 | 98 | **10.78** |
| **7** | 97 | 98 | 97 | **10.73** |
| | 95 | 100 | 97 | **10.69** |
| | 97 | 100 | 85 | **10.67** |

# Results - Clustering

| Cluster | LANG | MATH | HIST | PC1 |
|---|---|---|---|---|
| | 88 | 100 | 48 | **13.23** |
| | 67 | 82 | 18 | **12.83** |
| **1** | 70 | 90 | 30 | **12.49** |
| | 62 | 73 | 13 | **12.35** |
| | 63 | 93 | 38 | **12.14** |
| | **67** | **2** | **93** | **13.49** |
| | **68** | **12** | **98** | **12.21** |
| **2** | **93** | **22** | **97** | **10.09** |
| | 90 | 18 | 93 | **9.64** |
| | 78 | 18 | 95 | **9.51** |
| | 28 | 68 | 67 | **12.66** |
| | 38 | 80 | 70 | **12.54** |
| **3** | 55 | 83 | 92 | **12.13** |
| | 80 | 100 | 97 | **11.77** |
| | 2 | 2 | 28 | **11.29** |
| | **13** | **20** | **98** | **17.76** |
| | **7** | **20** | **72** | **14.90** |
| **4** | **28** | **22** | **95** | **14.63** |
| | **38** | **27** | **97** | **13.32** |
| | 23 | 12 | 67 | **11.25** |

| Cluster | LANG | MATH | HIST | PC1 |
|---|---|---|---|---|
| | 55 | 97 | 50 | **12.87** |
| | 58 | 97 | 45 | **12.79** |
| **5** | 55 | 97 | 67 | **12.64** |
| | 18 | 58 | 7 | **12.52** |
| | 37 | 82 | 60 | **12.47** |
| | **78** | **65** | **23** | **9.74** |
| | **77** | **48** | **17** | **8.88** |
| **6** | **93** | **60** | **40** | **8.84** |
| | **92** | **50** | **33** | **8.63** |
| | 78 | 23 | 13 | **8.41** |
| | 98 | 100 | 100 | **12.12** |
| | 93 | 100 | 98 | **10.78** |
| **7** | 97 | 98 | 97 | **10.73** |
| | 95 | 100 | 97 | **10.69** |
| | 97 | 100 | 85 | **10.67** |

**(Scores labeled as worth to investigate)**

## Conclusions

- Outlier Detection methods + PCA enable the identification of suspicious **scores**
- **Clustering** methods allow us to interpret the groups according to the patterns
- Awaiting for the analysis of the last standardize PDT process

## Future Work

- Analyze outliers at the level of **test rooms** and **test centers**
- Detection of **odd score patterns** in public schools

# Conclusions - Future Work

**Conclusions**

- Outlier Detection methods + PCA enable the identification of suspicious **scores**
- **Clustering** methods allow us to interpret the groups according to the patterns
- Awaiting for the analysis of the last standardize PDT process

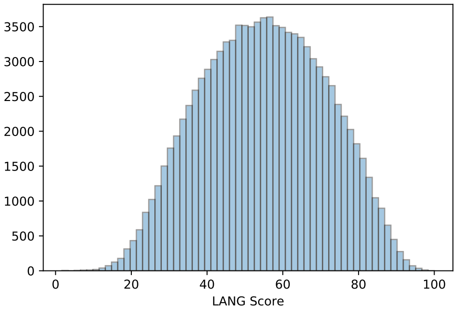
**Future Work**

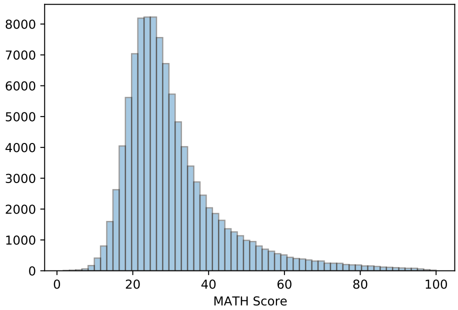
- Analyze outliers at the level of **test rooms** and **test centers**
- Detection of **odd score patterns** in public schools

# Questions

?

# Appendix: PDT Scores Distribution and Correlations




| | LANG | MATH | HIST |
|---|---|---|---|
| LANG | 1.00 | | |
| MATH | 0.56 | 1.00 | |
| HIST | 0.74 | 0.61 | 1.00 |

Table: LANG/MATH/HIST test scores correlation

# Appendix: Outlier Detection Methods Scoring

1. Mahalanobis Distance:

$$d_M(X^{(i)}, \hat{\mu}) = \sqrt{(X^{(i)} - \hat{\mu})^\top \widehat{\Sigma}^{-1} (X^{(i)} - \hat{\mu})}$$

2. Robust Mahalanobis Distance:

$$\widehat{\Sigma}_R = \sum_{i=1}^{n} \frac{K(d_M(X^{(i)}, \widetilde{X})^2)(X^{(i)} - \widetilde{X})(X^{(i)} - \widetilde{X})^\top}{\sum_{j=1}^{n} K(d_M(X^{(j)}, \widetilde{X})^2)}$$

where $K(u) = e^{-0.1u}$, and the distance then is computed as

$$d_{RM}(X^{(i)}, \widetilde{X}) = \sqrt{(X^{(i)} - \widetilde{X})^\top \widehat{\Sigma}_R^{-1} (X^{(i)} - \widetilde{X})}$$

3. iForest, and Extended iForest:

$$s(X^{(i)}, n) = 2^{-\frac{\mathbb{E}(h(X^{(i)}))}{c(n)}}$$

where $h()$ denotes the depth of the tree in which the $i$-th point is left as an outlier. $c(n)$ s the average path length of unsuccessful search in a Binary Search Tree with $n$ elements.

# Appendix: Anomaly Scores Correlation

Table: Correlation Matrix with Spearman's $\rho$ coefficient

| | Mahala. | R. Mahal. | iFor. | E. iFor | PC1 |
|---|---|---|---|---|---|
| Mahalanobis | 1.00 | | | | |
| R. Mahalanobis | 0.99 | 1.00 | | | |
| iForest | 0.98 | 0.97 | 1.00 | | |
| E. iForest | 0.97 | 0.98 | 0.98 | 1.00 | |
| PC1 | 0.99 | 0.99 | 0.99 | 0.99 | 1.00 |