MEIA -2025/2026

Project 1

Machines Dataset Analysis

Project done by:

IST No	Name	Email
87559	Pedro Gonçalves	pedrorenato@tecnico.ulisboa.pt
103629	Henrique Miguel Prates Santos	henrique.m.p.santos@tecnico.ulisboa.pt
106100	Guilherme Bertrand Melo de Sousa	guilherme.b.melo.de.sousa@tecnico.ulisboa.pt
106156	Daniel José Pinheiro da Costa	daniel.p.da.costa@tecnico.ulisboa.pt
106207	Maria Beatriz Mimoso Teles	maria.teles@tecnico.ulisboa.pt
107558	Ana Daniela Carona da Silva	daniela.c.silva@tecnico.ulisboa.pt

Introduction

This study provides an overview of the analysis conducted on a subset of the **machines** data set from the rrcov package, using **exploratory statistics and PCA** to understand its structure and variable relationships. It also outlines the motivation for examining how an introduced outlier affects both classical and robust PCA approaches.

Goals of the Study

- Explore the data set using statistical summaries and graphical methods.
- Apply PCA on original-scale and standardized variables to compare dimensionality reduction.
- Identify which PCA retains ≥95% variance and interpret the selected components.
- Assess the effect of an introduced outlier on classical PCA and compare it with robust MCD-based PCA.

1. Exploratory Data Analysis

We start by exploring the dataset to understand how the variables behave, how they differ in scale, and how they relate to one another. This initial exploration is important because it helps identify dominant variables, potential outliers, and redundant information that may influence later methods such as PCA.

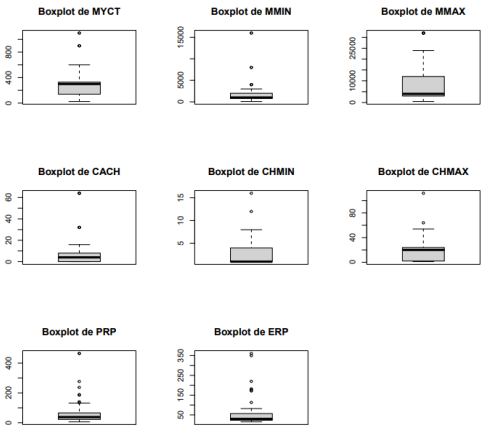
To do this, we examined summary statistics and four graphical tools: boxplots, a correlation heatmap, a histogram, and a scatterplot.

The summary statistics show clear differences between variable groups. The memory variables **MMAX** and **MMIN** have much larger numerical values and a much wider range compared to the remaining variables. They display strong variability, with some values far from the majority of observations. In contrast, variables such as **CACHMAX**, **CACHMIN**, **ERP**, and **PRP** are more compact and show less dispersion.

Because of these differences in scale and variability, these memory variables are expected to play a dominant role in any method based on variance, including PCA.

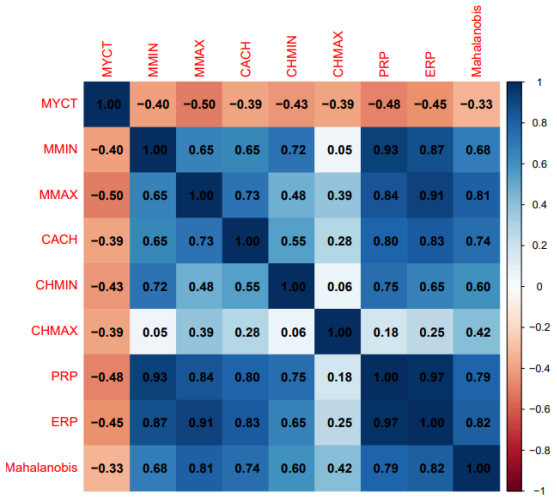
Graphical Analysis:

Figure 1 – Boxplots



The boxplots illustrate the distribution of each variable. As expected, **MMAX** and **MMIN** show the widest spread and several unusually large values, which appear as potential outliers. The remaining variables have tighter ranges, suggesting more consistent values across observations. This reinforces the idea that not all variables contribute equally to the total variability in the dataset.

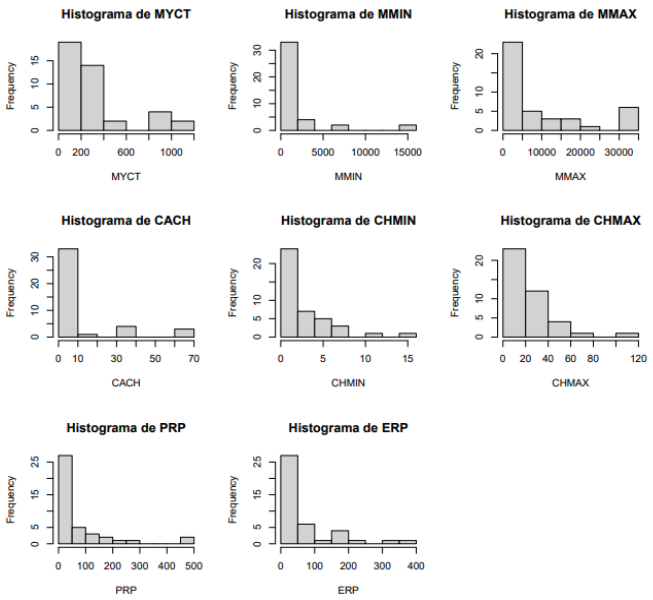
Figure 2 – Correlation Heatmap



The heatmap shows the pairwise correlations between variables. The strongest pattern is the very high correlation between **MMAX** and **MMIN**, meaning machines with high maximum memory also tend to have high minimum memory. Performance-related variables (**ERP** and **PRP**) display moderate correlations with each other. Cache variables show weaker relationships.

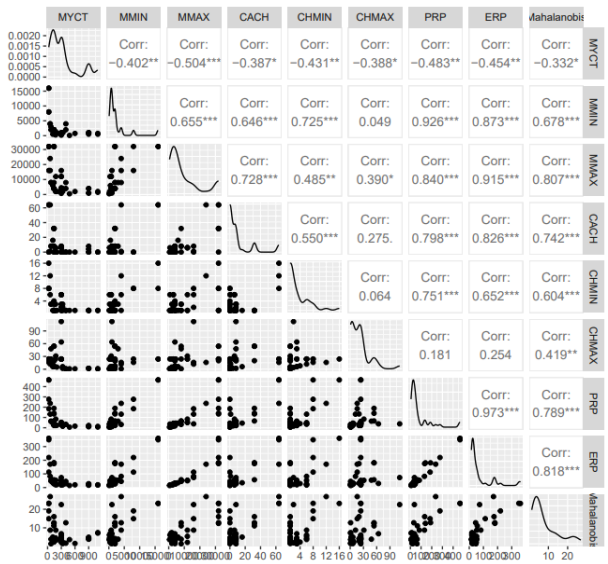
Overall, several variables share similar information, indicating that the dataset includes redundant features that may not all be necessary for dimensionality-reduction tasks.

Figure 3 – Histograms



The histogram focuses on the distribution of one selected variable (e.g., **MMAX**). The shape is clearly skewed: most machines have relatively low to medium values, while a few machines have extremely high memory capacity. These extreme values are visually highlighted and may influence classical statistical methods.

Figure 4 – Scatterplots

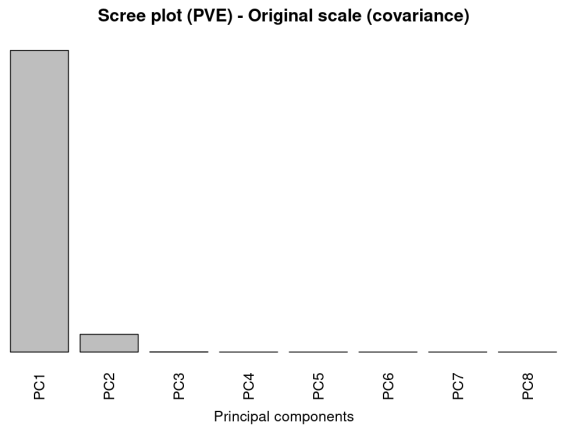


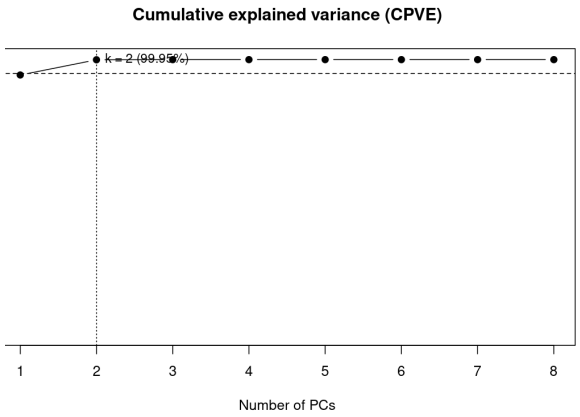
The scatterplot shows the relationship between two variables (e.g., **MMAX** and **MMIN**). Most points are concentrated in a compact region, confirming that typical machines share similar characteristics. A few observations lie far from this cluster, suggesting machines with unusually high specifications. These points may act as influential observations in downstream analysis.

2. PCA with Original Scale Variables

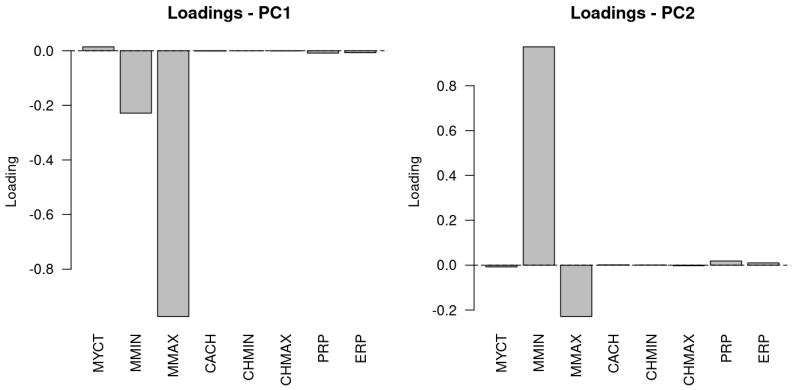
PCA was applied to the 8 variables in their original units (covariance-based PCA). Variables were mean-centered (no standardization), so higher-variance variables can dominate the first components.

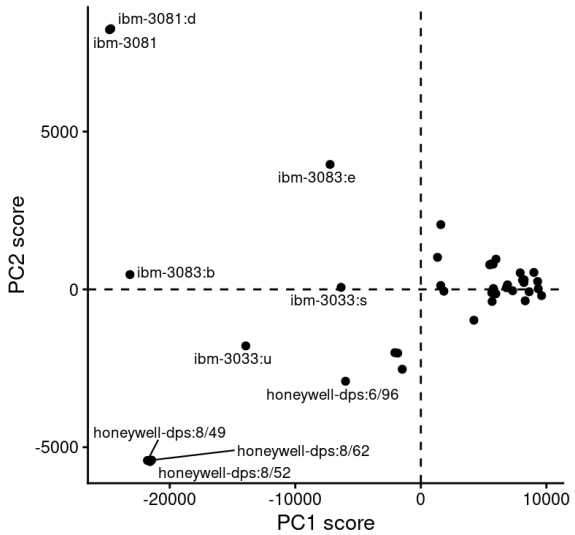
Variance explained: PC1 explains 94.41% and PC2 explains 5.53%; PCs 3–8 are negligible. Cumulative variance is 0.944 after PC1 and 0.99945 after PC2, so the minimum number of components to keep at least 95% variance is k=2 (≈99.95% retained). This **reduces the data from 8 dimensions to 2**.





Interpretation of retained PC's: PC1 is driven mainly by **MMAX** (dominant) and **MMIN** (secondary), representing an overall memory-capacity direction. PC2 is driven mostly by MMIN with an opposite contribution from MMAX, capturing a minimum-memory contrast not explained by PC1.



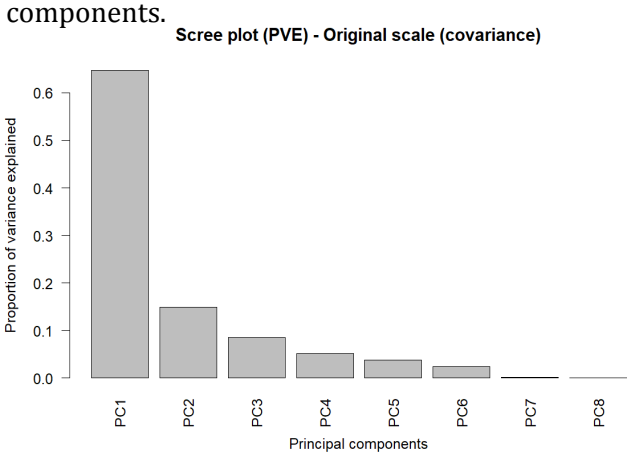


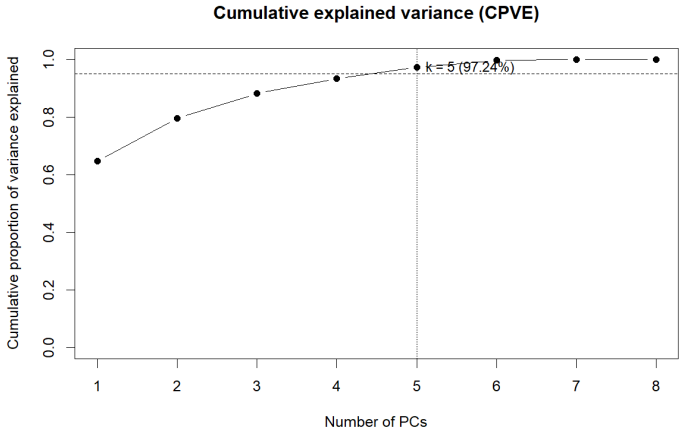
Scores Structure: Most observations cluster near the origin, with a few extremes mainly along PC1 which is consistent with memory variables explaining most variation.

3. PCA with Standardized Variables

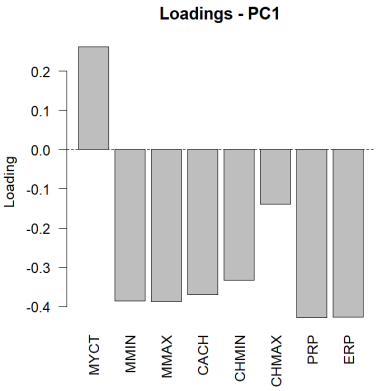
PCA was applied to the 8 standardized variables, using the classical correlation matrix. All variables were transformed, ensuring that they contribute equally to the analysis and preventing higher-variance variables from dominating the principal components.

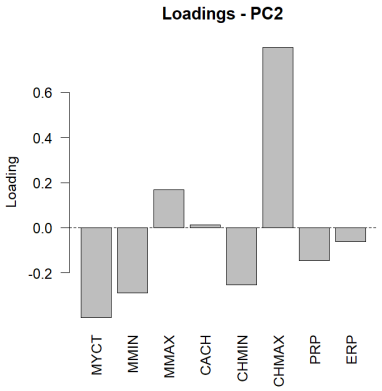
Variance explained: The scree plot shows that PC1 explains about 65% of the variance, PC2 adds around 15%, and PC3 contributes roughly 9%. The remaining components each explain very little. The cumulative variance plot indicates that: PC1–PC3 capture about 90% of the total variance. To retain at least 95%, four components are needed. The marked threshold on the plot shows that five components explain about 97.24% of the variance. The dimensionality of the data can be reduced from 8 variables to 5 principal components.

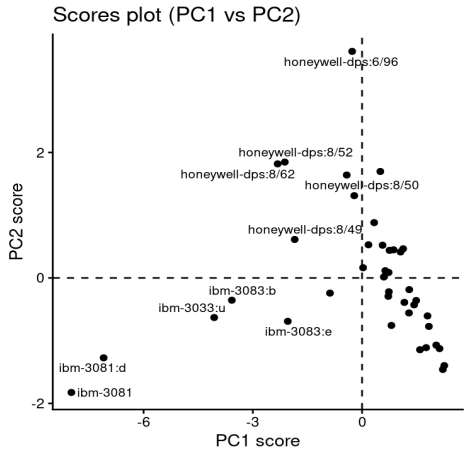




Interpretation of retained PC's: PC1 is the dominant dimension separating observations, consistent with the much higher variance explained. Only a few observations deviate strongly from the main cluster, likely due to extremes in the variables affecting PC1 (e.g., very high or very low values across several measurements). Variation along PC2 mainly reflects differences driven by CHMAX, but this is a weaker and more localized effect.







Scores structure: Most observations cluster near the positive PC1 region, with only a few strong outliers, mainly separated along PC1. Negative scores indicate values below the overall mean, while positive scores reflect values above it. The IBM-3081 systems appear far left because several of their variables are well below the mean, whereas the Honeywell-DPS models lie on the upper right because they have above-average values on the variables that most influence PC1 and PC2. Points near the center are close to the mean across all variables.

4. Comparison and Decision (≥95% Variance)

For the purpose of dimension reduction while retaining at least 95% of the total variance, we recommend the **classical PCA on the original scale** (covariance matrix).

Original-scale PCA (covariance):

PC1 explains ≈ 94%, and PC1+PC2 ≈ 99.95% → 2 PCs are sufficient to exceed 95%.

Standardized PCA (correlation / z-scores):

The cumulative curve reaches ≥95% only at about k = 5 (≈ 97.24%) → **5 PCs are required.**
Since the goal is to reduce dimensionality as much as possible while keeping ≥95% variance, **2 PCs (covariance PCA) is clearly preferable to 5 PCs** (standardized PCA).

Interpretation of retained principal components (covariance PCA: PC1 and PC2)

As we've seen before, based on the loadings from the original-scale PCA:

- **PC1** (dominant component; ~94%) is essentially a memory capacity axis, driven mainly by **MMAX** (and to a lesser extent **MMIN**).
  Observations separate strongly along PC1 primarily due to differences in maximum memory.

- **PC2** (adds the remaining variance to reach ~99.95%) is mostly a minimum memory axis, dominated by **MMIN**, with **MMAX** contributing in the opposite direction.
  This component refines separation based on minimum memory and its contrast with maximum memory.

Extremes in the PC1–PC2 space

The most extreme systems (ibm-3081, ibm-3081:d, ibm-3083:b and honeywell-dps:8/49, :8/52, :8/62) all share very large maximum memory (MMAX = 32000), which explains their extreme scores on PC1, our "memory capacity" axis.

PC2 then separates these high-memory machines: the IBM 3081 models combine high MMAX with high MMIN and high PRP/ERP (high-end, consistently large-memory and fast systems), whereas the Honeywell DPS-8 machines have the same MMAX but much smaller MMIN, higher MYCT and lower PRP/ERP, with ibm-3083:b occupying an intermediate profile.

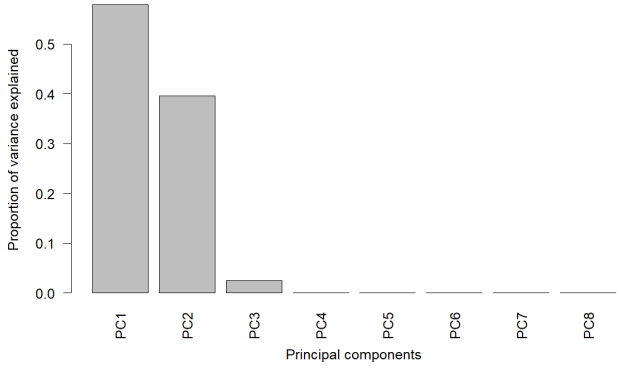
The original values of these variables for each extreme machine and their respective scores can be analysed in the **Annex** section as tables **4.1** and **4.2**.

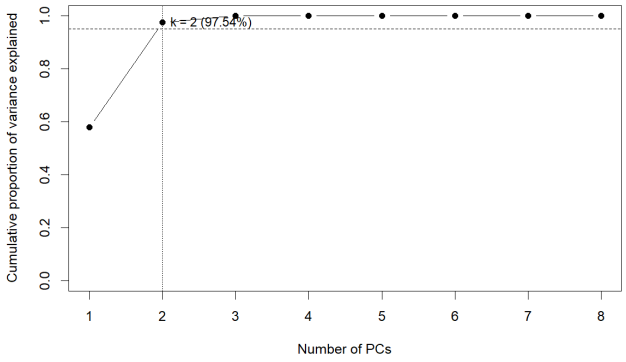
5. Outlier Introduction and Impact Analysis

An outlier of extreme values was introduced into row hp-3000/64 of the dataset, likely to simulate a gross measurement/error or an extreme machine configuration, showing how a single atypical case can influence multivariate summaries. The outlier remains in the raw, standardized, and original-scale analyses so its effect on classical and robust PCA methods can be directly compared.

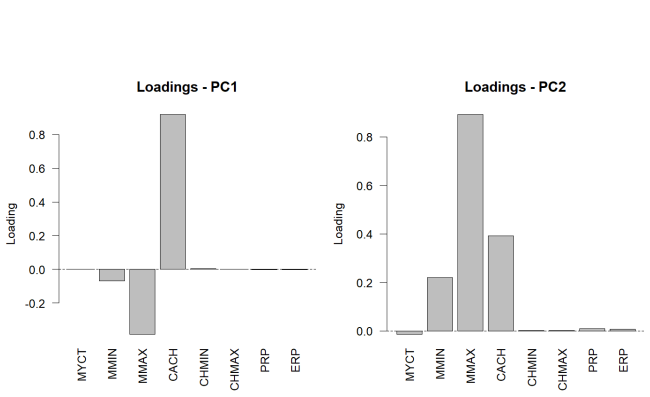
Classic PCA - Impact Analysis

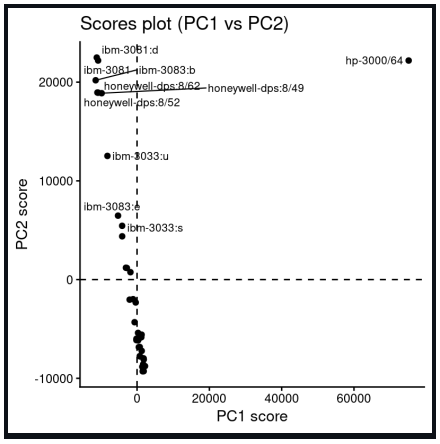
The injected observation has overwhelmingly altered the classical PCA summary. The first two components together explain almost all variance: the first component is dominated by the extreme variable (the 80000 entry in the inserted vector), and the scores plot shows a single extreme point far from the bulk while the remaining observations are compressed near the origin. In short, the retained PCs mainly describe the outlier direction rather than the internal structure of the majority of the data.





Considering the Proportion of variance PC1 and PC2 capture the vast majority of variance, which indicates a single dominant direction of variability. The CPVE curve reaches the 95% threshold at k = 2, meaning the routine will retain only the two PCs dominated by the outlier-driven variance rather than a balanced, multi-variable structure across the dataset.





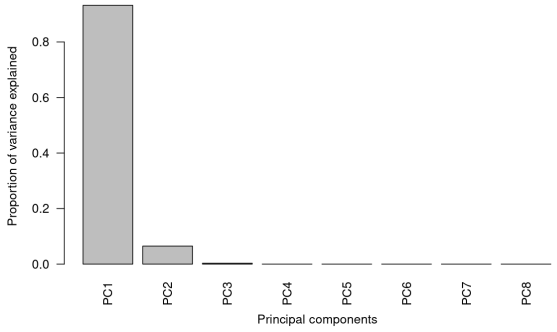
One variable has a very large positive loading on PC1 while most other variables have loadings that are negative or near zero, meaning it captures the outlier's direction. PC2 also emphasized variables tied to the outlier and most others had near-zero loadings. Instead of the real, recurring patterns among machines they showed originally, the outlier loadings tell you mainly about that one extreme case which is misleading.

The scores plot shows one observation with a massive positive PC1 score far to the right, unlike the others all near PC1=0. This confirms the outlier dominates the PC1 direction and distorts the classical PCA geometry.

Robust PCA - Impact Analysis

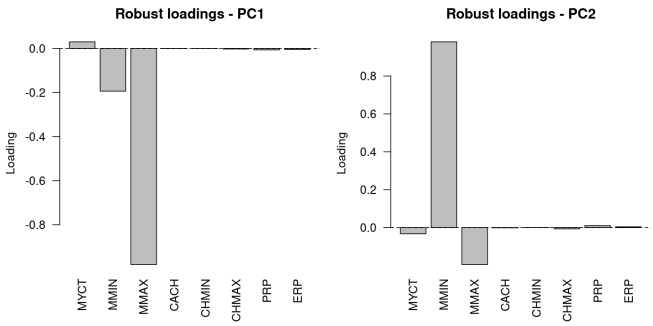
Robust PCA with **MCD** (Minimum Covariance Determinant) first finds a "clean" subset of the data (here **α = 0.75**, so ~75% of the machines) whose covariance has the smallest determinant. PCA is then built from that robust center/covariance. The key consequence is simple: **the outlier does not get to "steer" the principal directions**.

From the MCD computation output:



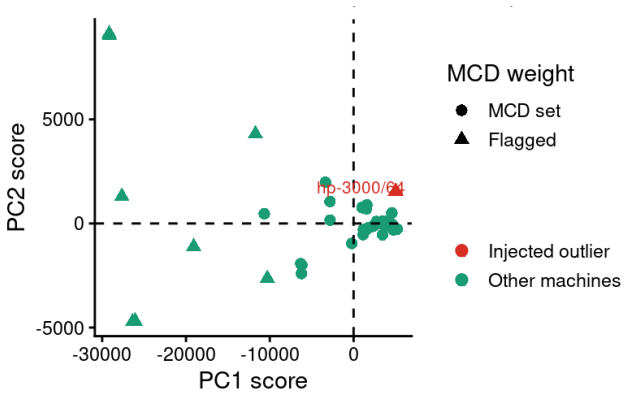
- **Variance explained (robust):**
  PC1 ≈ **93.19%**
  PC2 ≈ **6.50%**
  By **2 PCs** we already reach **99.69%** cumulative variance.
  Interpretation: once the outlier's leverage is removed, the dataset is essentially **almost 1-dimensional**, with a small second mode.

- **Robust loadings (what PC1/PC2 mean):**



  PC1 is dominated by MMAX (loading ≈ −0.9806), with a smaller contribution from **MMIN** (≈ −0.1937).
  PC2 is dominated by MMIN (≈ +0.9802), with smaller MMAX (≈ −0.1947).
  Interpretation: robust PCA says the main structure is driven by **memory size** variables (MMAX/MMIN), not by the injected extreme configuration.

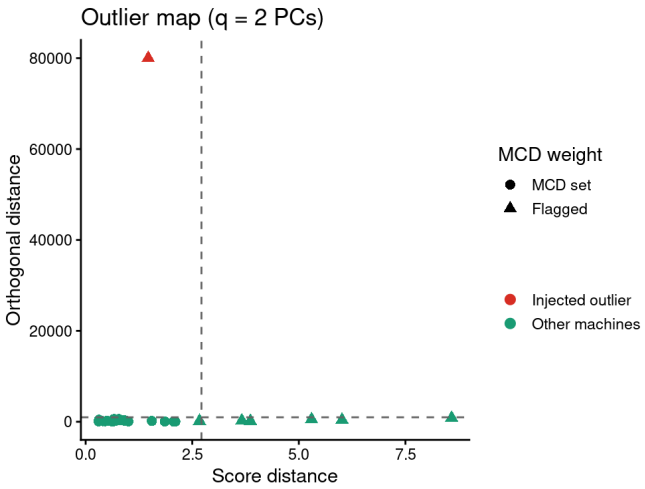
- **Scores + outlier handling:**
  In the robust scores plot, **hp-3000/64 is not thrown far away along PC1/PC2** (because it's downweighted), but in the **outlier map** it shows up clearly by having a **very large orthogonal distance**: it doesn't fit the low-dimensional PCA model built from the "good" subset.
  Interpretation: MCD separates *"where it lies in the*



model space" (score distance) from *"how badly it violates the model"* (orthogonal distance). That's exactly what we want for anomaly detection.

Outlier in PCA classic and Robust - impact analysis comparison

The classical plots show the typical failure mode: **one atypical observation dominates the covariance**, which rotates the PCs and stretches the score space.



1) Scree / variance explained

Classical: PC1 ≈ **0.58**, PC2 ≈ **0.40** → variance looks "split" across two PCs.
Robust (MCD): PC1 ≈ **0.93**, PC2 ≈ **0.065** → variance is mostly one direction.
Interpretation: the outlier makes classical PCA look more 2D than it really is.

2) Score plot (PC1 vs PC2)

Classical: *hp-3000/64* sits extremely far to the right (huge PC1), forcing the axis scale to expand so much that **most machines get squashed near the origin**.
Robust: the main cloud is readable; the outlier doesn't "pull" the axes, and flagged points are visible without destroying the scale.
Interpretation: classical PCA loses visual/analytic resolution for the majority because the outlier dictates the geometry.

3) Loadings / biplot interpretation

Classical: PC1 is heavily aligned with **CACH** (big positive loading), meaning the injected point has effectively redefined what "most variance" means.
Robust (MCD output): PC1/PC2 are driven mainly by **MMAX/MMIN**, which is much more consistent with the "typical" machines' structure.
Interpretation: classical PCA can give a misleading story ("cache drives everything") because the outlier reorients the components.

Bottom line:
With one gross outlier, **classical PCA changes the coordinate system to explain that one point**, while **robust MCD PCA keeps the coordinate system representative of the majority**, and still flags the outlier via poor model fit (large orthogonal distance).

Conclusions

This project used exploratory analysis and PCA on the machines dataset to identify the main patterns of variability and the impact of scaling and outliers. Memory features (MMAX and MMIN) showed the greatest dispersion and strong correlation, so they dominated variance-based structure. Covariance PCA reduced the data most effectively (about 2 PCs to exceed 95% variance), while correlation PCA needed more components (about 5 PCs). Adding an extreme outlier distorted classical PCA, whereas robust MCD-PCA limited its influence and better preserved the underlying structure.

Annex

4.1

Machine	MMAX	MMIN	MYCT	PRP	ERP
ibm-3081	32000	16000	26	465	361
ibm-3081:d	32000	16000	26	465	350
ibm-3083:b	32000	8000	26	277	220
honeywell-dps:8/62	32000	2000	140	189	181
honeywell-dps:8/52	32000	2000	140	141	181
honeywell-dps:8/49	32000	2000	140	134	175

4.2

Machine	PC1	PC2	Distance
ibm-3081	-24707.84	8133.775	31531.31
ibm-3081:d	-24707.76	8133.6596	31531.2
ibm-3083:b	-22876.12	343.3633	28632.63
honeywell-dps:8/62	-21501.79	-5498.6492	27803.82
honeywell-dps:8/52	-21501.37	-5499.5427	27803.58
honeywell-dps:8/49	-21501.24	-5499.6479	27803.47