Chat GPT-4o: A powerful Tool to Quickly Identify Anomalous Lots in a Dataset

1. INTRODUCTION

As highlighted in the first post of this blog (05/09/2018), pharmaceutical Quality Control (QC) data is generally summarized by organizing it into data tables (or data matrices), each row of which contains the results of several measurements (*e.g.*, assay, pH value, *etc.*) carried out on a specific batch. From a QC perspective, each data table row provides the "analytical profile" of that specific batch while the entire data table provides, in a sense, the "analytical fingerprint" of that manufacturing process.

Regulations require these datasets to be evaluated for knowledge and insight, which can be performed using a univariate (conventional) or a multivariate approach.
Multivariate data analysis and visualization can easily provide an overview of the whole manufacturing process and reveal potentially anomalous lots.

In the first post of this series, such an analysis was conducted using a specific program in R that returned data tables and graphs. In this post, I want to show how, using the same initial dataset, it is possible to obtain the same results in a few seconds without having to write any program in R, simply by asking Chat GPT-4o.

2. EXPERIMENTAL SECTION

As a case study, the same data set used in the first post of this series six years ago was considered (see Table 1). It concerns the analytical results of thirty-one (31) production batches of a hypothetical active ingredient. All batches were produced using the same manufacturing method. The dataset was uploaded as an .xlsx file.

Table 1: Analytical Results of 31 Production Batches

lot	h20	ph	assay	sm	known	unk	total	solv1	solv2	solv3
1	1.80	5.7	86.6	0.0016	0.02	0.05	0.3	3.3	2.5	0.01
2	1.40	5.7	86.9	0.0016	0.02	0.04	0.3	2.2	1.8	0.01
3	1.80	5.8	88.6	0.0016	0.02	0.04	0.2	2.3	2.1	0.01
4	1.60	5.5	86.7	0.03	0.02	0.04	0.3	3.4	2.8	0.01
5	2.90	6.8	87.9	0.0016	0.02	0.05	0.2	1.7	2.8	0.06
6	1.80	6.9	90.2	0.0016	0.02	0.05	0.2	0.9	1.5	0.04
7	2.10	6.8	88.3	0.0016	0.02	0.04	0.2	1.9	2.8	0.06
8	3.00	5.6	90.6	0.0016	0.02	0.04	0.1	1.1	1.9	0.03
9	2.70	5.7	89.8	0.0016	0.01	0.04	0.1	1.7	2.8	0.08
10	2.30	5.6	89.9	0.0016	0.01	0.04	0.1	0.9	1.6	0.05
11	1.90	5.3	90	0.0016	0.02	0.04	0.2	1.2	2	0.03
12	2.20	5.8	89.6	0.01	0.02	0.05	0.2	2	2.5	0.75
13	2.00	8	88.1	0.02	0.01	0.04	0.2	0.4	0.8	0.03
14	2.00	8	88.1	0.02	0.01	0.04	0.2	1.3	2	0.05
15	1.60	6.8	89.2	0.01	0.01	0.06	0.2	1.1	2.1	0.07
16	1.70	6	89.6	0.01	0.02	0.14	0.3	1.3	2	0.1
17	1.70	7.7	91.5	0.02	0.02	0.06	0.2	1.2	1.4	0.04
18	1.60	6	90.9	0.01	0.02	0.04	0.3	1.5	2.4	0.09
19	1.90	5.6	89.6	0.02	0.02	0.05	0.3	2	2.4	0.06
20	2.50	5.7	86.1	0.0016	0.01	0.1	0.2	3.3	3.9	0.1
21	2.50	5.5	90.7	0.0016	0.01	0.1	0.2	1.3	1.7	0.03
22	2.40	5.5	87.6	0.0016	0.01	0.1	0.2	1.2	1.6	0.04
23	2.40	5.6	87.9	0.0016	0.02	0.1	0.2	2.3	2.6	0.05
24	2.80	5.7	91	0.0016	0.01	0.1	0.2	1.4	1.8	0.04
25	4.10	5.3	87.6	0.0016	0.01	0.1	0.3	1.3	1.5	0.02
26	2.50	5.4	88.1	0.0016	0.04	0.2	0.6	0.8	0.9	0.01
27	3.60	5.4	83.8	0.0016	0.01	0.1	0.3	1.7	1.7	0.05
28	2.60	5.5	81.7	0.0016	0.0014	0.1	0.2	1.6	1.5	0.05
29	1.90	5.5	86.8	0.0016	0.0014	0.2	0.2	2.1	2.1	0.04
30	2.70	6	83.4	0.0016	0.014	0.1	0.1	4.8	4.5	0.07
31	1.90	5.6	88	0.0016	0.015	0.1	0.2	1.8	2.4	0.04

Table 2: Abbreviations and Specification Limits

Analytical parameter (or variable)	Units	Allowed Range of Variability	Analytical Technique	Abbreviation
pH	pH units	5.0 – 8.0	pH-metry	ph
Residual water content	%	1.0 – 5.0	Karl-Fisher titration	h2o
Assay	%	80 - 92	HPLC	assay
Starting material residual content	%	≤ 0.20	HPLC	sm
Largest known impurity	%	≤ 0.20	HPLC	known
Largest unknown impurity	%	≤ 0.20	HPLC	unk
Total impurities content	%	≤ 1.0	HPLC	total
Residual solvent 1 content	%	≤ 5.0%	Gas-chromatography	solv1
Residual solvent 2 content	%	≤ 5.0%	Gas-chromatography	solv2
Residual solvent 3 content	%	≤ 1.0%	Gas-chromatography	solv3

3. RESULTS

USER QUESTION 1:

The attached dataset refers to 31 batches of a given active pharmaceutical ingredient.

It contains, for each batch, the values of the analytical parameters reported on the release analysis certificate.

Using Principal Component Analysis can you build two-dimensional density maps and identify those lots that behave anomalously or differently from the others?

CHATGPT-4O ANSWER 1:

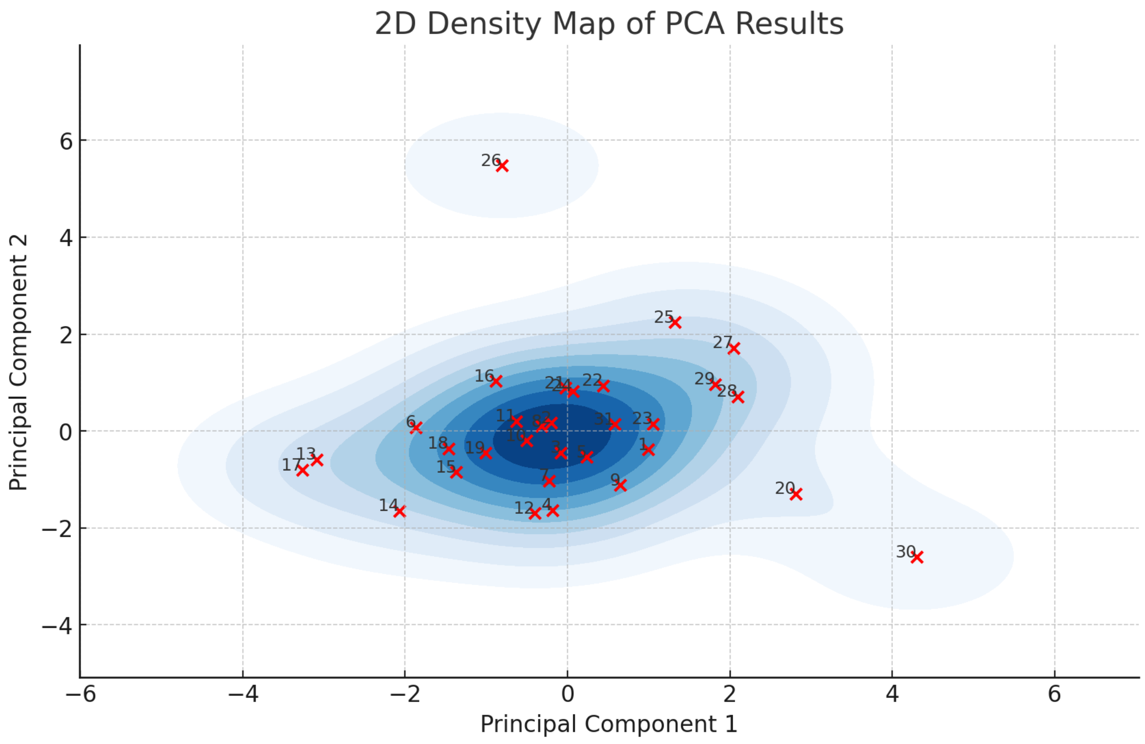
The dataset contains the following columns for 31 batches of a given active pharmaceutical ingredient:

- Lot: Lot number
- h20: Water content
- ph: pH value
- assay: Assay value
- sm: Specific measurement (*e.g.*, impurity level)
- known: Known impurities

- unk: Unknown impurities
- total: Total impurities
- solv1: Solvent 1 residue
- solv2: Solvent 2 residue
- solv3: Solvent 3 residue

To identify anomalous or differently behaving lots using Principal Component Analysis (PCA), we can perform the following steps:

1. Standardize the data.
2. Perform PCA.
3. Visualize the results using a two-dimensional density map.
4. Identify outliers.



The 2D density map of the PCA results is displayed. Each red dot represents a lot, and the density map shows the distribution of the lots in the space defined by the first two principal components. The lot numbers are also labeled for easy identification.

Interpretation

- **Clusters**: If you observe clusters of points, it means those lots have similar characteristics.
- **Outliers**: Points that lie far from the dense areas (clusters) can be considered as anomalous or behaving differently.

Next Steps

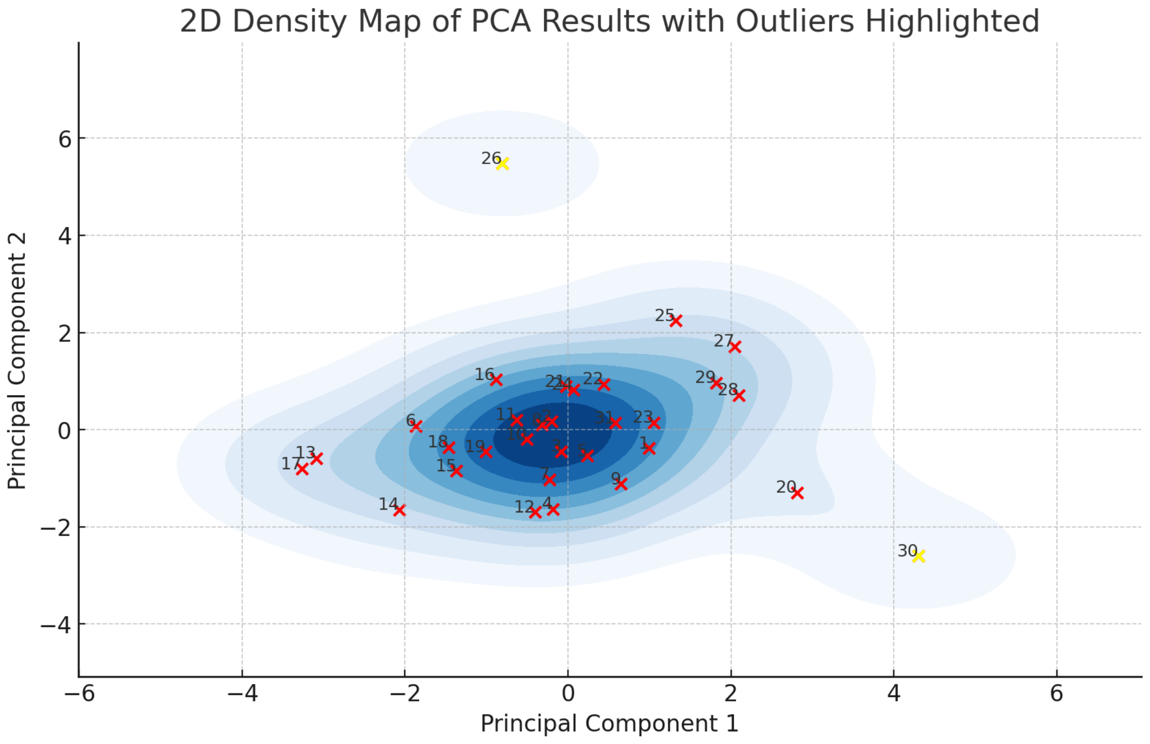
To identify the anomalous lots more clearly, we can use statistical methods like Mahalanobis distance or simply highlight those points that are significantly far from the center of the main cluster.

Would you like to proceed with identifying these outliers using a specific method, or is there any other specific analysis you would like to conduct on this data?

USER QUESTION 2:

Proceed identifying outliers, i.e., those points that are significantly far from the center of the main cluster.

CHATGPT-4O ANSWER 2:



The outliers have been identified and highlighted in yellow on the 2D density map. The table below shows the lots that were determined to be outliers based on their Mahalanobis distance:

Principal Component 1	Principal Component 2	Lot	Mahalanobis Distance
-0.810351	5.480502	26	3.765034
4.299214	-2.599267	30	3.190491

These lots are significantly different from the main cluster of lots, indicating potential anomalies or different behaviors.

If you need further analysis or another type of visualization, please let me know!

USER QUESTION 3:

What differentiates lots 26 and 30 from the others?

CHATGPT-4O ANSWER 3:

The comparison table highlights the differences between the outlier lots (26 and 30) and the non-outlier lots based on their z-scores. Here's a summary of the key differences:

- **Lot 26**
  - Water content (h20): Slightly higher than average (z-score = 0.43)
  - pH (ph): Lower than average (z-score = -0.78)
  - Known impurities (known): Significantly higher than average (z-score = 4.19)
  - Unknown impurities (unk): Significantly higher than average (z-score = 3.34)
  - Total impurities (total): Significantly higher than average (z-score = 6.36)
  - Solvent 1 (solv1): Lower than average (z-score = -1.25)
  - Solvent 2 (solv2): Lower than average (z-score = -1.98)

- **Lot 30**
  - Water content (h20): Slightly higher than average (z-score = 0.75)
  - pH (ph): Close to the average (z-score = -0.03)
  - Assay: Significantly lower than average (z-score = -2.30)
  - Known impurities (known): Close to the average (z-score = -0.12)
  - Unknown impurities (unk): Higher than average (z-score = 0.76)

- Total impurities (total): Lower than average (z-score = -1.95)
- Solvent 1 (solv1): Significantly higher than average (z-score = 4.29)
- Solvent 2 (solv2): Significantly higher than average (z-score = 3.94)

Summary:

- **Lot 26** is characterized by significantly higher known and unknown impurities, and total impurities compared to other lots. It also has slightly higher water content and lower solvent residues.
- **Lot 30** stands out due to its significantly lower assay value and higher solvent residues, while its impurity levels are more moderate compared to Lot 26.

These significant deviations suggest that lots 26 and 30 have different behaviors, potentially due to variations in the manufacturing process or raw material quality.

4. CONCLUSIONS

The analysis conducted using ChatGPT-4o demonstrates the powerful capabilities of AI in quickly identifying and interpreting anomalies within complex datasets. By asking a series of well-constructed, precise questions, ChatGPT-4o can efficiently summarize and visualize the information contained in data tables, making it accessible even to individuals who may not be experts in statistics or multivariate analysis.

This process eliminates the need for extensive programming by the data analyst, which was required for similar analyses conducted six years ago. The AI not only identifies anomalous batches but also provides insights into the reasons behind these anomalies, such as higher impurity levels or variations in solvent residues. Moreover, the AI hypothesized potential issues in the manufacturing process or raw material quality, based on significant deviations observed in certain lots. For example, the distinct behaviors of lots 26 and 30 suggested possible variations in the production process or differences in raw material quality.

However, it is essential to emphasize that knowledge of statistics remains crucial. To ask detailed and meaningful questions to the AI and to correctly interpret its answers, one must have a solid understanding of statistical concepts and methods. This foundational knowledge ensures that the insights gained from AI analyses are accurate and actionable.

ChatGPT-4o proves to be an invaluable tool for the rapid and effective review of data tables, such as those for Annual Product Quality Reviews (APQR). It helps in pinpointing anomalies, understanding their nature, and ultimately enhancing the knowledge of the production process and the quality of the final product. Nonetheless, the role of a knowledgeable statistician is irreplaceable in guiding the AI and making informed decisions based on its outputs.

R. Bonfichi © 2024. All rights reserved.