

Data science assistant LLM

May 22, 2025

Natasja Vitoft **Lasse Baggesgård Hansen**
cph-nn194@cphbusiness.dk **cph-lh479@cphbusiness.dk**

1 Introduction

For this project we tried to create a system for feeding context to an LLM, making it better suited for assisting with data analysis tasks.

Our goal was to make it possible to upload datasets in CSV format to the system, which then could be formatted into a prompt template. The LLM should then be able to understand the statistics and correlations of the dataset represented in CSV. We hope this will make it more suited to aid the analysis process by always having a detailed context of the users data when asked questions.

We set up a chat system using Streamlit, where the user can upload their dataset. Descriptive statistics and correlation matrices are then calculated using the Pandas Python library, and inserted into the user prompt, as a template together with a 5 row sample of the the dataset, before being sent as an API request to the LLM.

The user can then keep this context intact through their whole chat session, optionally providing additional context directly in the prompt.

1.1 Optimizing machine learning models

We want to use the AI as an “Extra pair of eyes”, for optimizing our data cleaning and model tuning. for this we formulated the following questions:

1. “I have the following dataset. What would be the best strategy for predicting the continuous numeric feature Total.Cup.Points?”
2. “We have trained a Random Forest Regression model to predict the Total.Cup.Points variable using this dataset. Can you give us ideas on how we can tune our model to perform better on our dataset, without overfitting the data?”
3. “We’re getting the following r^2 score and MSE score Using a random forest regressor.

—Score table—

What does these metrics tell us about how well the data fits the model?”

1.2 Interpreting results:

We are interested in finding out if we can gather additional insights from the results of our models by the following questions:

4. From the dataset, we ran a PCA on the features:

”Aroma”, ”Flavor”, ”Acidity”, ”Body”, ”Balance”, ”Aftertaste”, ”Uniformity”, ”Sweetness”. Which yielded 3 principal components. Their weights are seen here:

—PC weight table—

We then ran a DBSCAN clustering analysis on the principal components, which generated 3 clusters, that seem to split from each other primarily on the axis of PC3 and PC2, while correlating negatively along the axis of PC1.

What insights do you believe we can potentially gather from this?

2 Process

We started using DeepSeek-R1:1.5b because of its reputation as a good compact LLM with chain-of-thought capabilities. The model hallucinated a lot, and was not really giving any valuable output. We tried formatting the prompt template with detailed descriptions of the given data, but it did not seem to notice it at all. We tried upgrading the model to DeepSeek-R1:7b, which significantly lowered the amount of hallucinations on the cost of longer execution times. The model still did not seem to notice the provided context.

At this point we noticed from Ollama output logs, that our input was being truncated due to input token limitations. After updating our API calls with tripled limits, the LLM finally started to notice the context. However, it was still not quite capable of making sense of it. When asked about what it could make of the data, it often answered very speculatively, like it didn't know what relations the numbers had.

We did some research into other models, and was recommended Qwen3, as it should be better suited for handling tabular data. We chose Qwen3:8b as it was similar in size to The previous DeepSeek model used.

This time the model reacted with a much better understanding of the provided context, and was capable of answering questions about their function, and reason about the meaning of data provided.

The actual question 4 prompt sent, and the response received from Qwen3 can be found in Appendix section 5

We received decent valuable advice from the model, and could apply further consideration to analysis, but it didn't add directly applicable adjustments to the analysis.

3 Comparative analysis

In this comparative analysis we will compare Qwen, as the local AI solution and ChatGPT. We will compare the two language models based on different metrics such as performance/efficiency and security/privacy.

After having compared the Qwen model and ChatGPT we see that the response time when prompting the Qwen model is much higher than when prompting ChatGPT, as we expected. When prompting the Qwen model we usually had to wait around 10 minutes, more or less, and with ChatGPT we received an answer almost immediately. So when comparing the two models in terms of response time, ChatGPT is clearly much faster. We did however use our private machines for the comparison which doesn't have the necessary CPUs or GPUs to run the local models as efficiently as possible.

We prepared four prompts to give to our Local AI solution and afterwards the same prompts to ChatGPT. It took a very long time to get all of our answers from the local AI solution but it was very fast with ChatGPT. We did however experience that the local AI solution sometimes performed better with accuracy and interpretation than ChatGPT. When prompting ChatGPT there were a few times where it didn't seem to understand the question as well as the Qwen model. It could seem that ChatGPT might be under the same input token limitations that we previously experienced, and therefore cannot fit the whole context.

When we look at security/privacy it is generally known that local AI models are a better solution than cloud based models like ChatGPT. So when it comes to security and privacy the Qwen model clearly has an advantage because we can decide how our data is managed and it is only local on our machines. By using a model like ChatGPT, we don't have any influence on how the data is processed. We have to count on the company to avoid potential data leaks. In our case where we investigate coffee quality, a local AI model in terms of privacy and security might not be that important or necessary, because it's not really

sensitive data we are working with. However if our business case included data that were more sensitive, like sensitive personal data about people, a local AI could be a good solution to have more control in regards of data processing and maybe even necessary because of data legislation.

<https://research.aimultiple.com/cloud-llm/>

4 Conclusion

In conclusion, we believe that the system has proved itself slightly useful, but at a high cost.

The Qwen model gave us a some good condensed takeaways from our data analysis, more confidence in our regression models performance, and some interesting ideas on how to perceive the outcome of our clustering analysis.

The system can give decent advice on the provided datasets without the user having to write long elaborate prompts to explain everything. It does however, take around 10 minutes to give a response. The Ollama framework and model takes around 25Gb of disc space, and around 16Gb of ram while running, even with GPU utilization.

The privacy aspect might be worth it for companies willing to spend on dedicated hardware, but it doesn't seem feasible for a private person because of the high spec requirements, as opposed to online hosted models.

5 Appendix

5.1 Cluster insight prompt with template

"You are provided a user prompt and a context. Use the context as a basis for answering the user prompt as good as you can.

All dataframe information is provided in CSV format, but represents a Pandas Dataframe. Use this as a background for responding.

Only consider the context if it is provided explicitly

Context:

Act like you are:

You should be addressing:

Dataset Sample

If present, this is a 5 row sample of the users dataset:

—DATASET SAMPLE—

Dataset Descriptive Statistics

This is generated from the users dataframe using pandas df.describe(). The rows from top is: count, mean, standard deviation, 25% quartile, 50% quartile, 75% quartile, max:

—DATASET DESCRIBE—

Dataset Correlation Matrix.

This is generated from the users dataframe using pandas df.corr() The rows represent the columns. The values is the correlation between features:

—DATASET CORRELATION MATRIX—

User Prompt:

From the dataset, we ran a PCA on the features: "Aroma", "Flavor", "Acidity", "Body", "Balance", "Aftertaste", "Uniformity", "Sweetness". Which yielded 3 principal components. Their weights are seen here:

—PCA WITH WEIGHTS MARKDOWN—

We then ran a DBSCAN clustering analysis on the principal components, which generated 3 clusters, that seem to split from each other primarily on the axis of PC3 and PC2, while correlating negatively along the axis of PC1.

What insights do you believe we can potentially gather from this?

5.2 Response from local model Qwen3:8b

—THINKING—

Okay, let's try to figure out what insights we can get from the PCA and DBSCAN results. First, I remember that PCA reduces the dimensionality by combining features into principal components (PCs) that capture the most variance. The user mentioned three PCs, and the weights (loadings) show which original features contribute to each PC.

Looking at PC1, the highest loadings are for Aroma, Flavor, Acidity, Body, Balance, and Aftertaste. These are all sensory attributes related to the coffee's taste and mouthfeel. So PC1 probably represents a general quality or intensity of the coffee's flavor profile. High values in PC1 might indicate coffees with strong, balanced, and aromatic characteristics.

PC2 has high loadings on Uniformity and Sweetness, with some contributions from Body and Acidity. Uniformity relates to consistency in the cup, and Sweetness is a taste note. So PC2 might capture the sweetness and uniformity aspects, maybe indicating how smooth or consistent the coffee's taste is.

PC3 has high loadings on Uniformity and Sweetness again, but wait, looking at the weights, PC3's Uniformity and Sweetness have higher weights than PC2. Wait, no, PC2's Uniformity is 0.6593 and Sweetness 0.7332. PC3's Uniformity is 0.7315 and Sweetness 0.6741. So both PC2 and PC3 have high Uniformity and Sweetness. Maybe PC2 and PC3 are capturing different aspects of these features. Wait, maybe I need to check the correlation matrix to see if Uniformity and Sweetness are correlated. Wait, the correlation matrix shows that Uniformity and Sweetness have a correlation of 0.0083, which is almost zero. So they might be independent. So PC2 and PC3 might be capturing different dimensions of Uniformity and Sweetness. Wait, but the loadings for PC2 and PC3 for these features are similar. Maybe there's a mistake here, but assuming the data is correct, PC2 and PC3 could be representing different aspects. Alternatively, maybe the user's data has some noise, but I'll proceed with the given info.

The DBSCAN clustering split into 3 clusters based on PC3 and PC2, with negative correlation along PC1. So the clusters are separated along PC2 and PC3, which might represent different combinations of Uniformity and Sweetness. The negative correlation along PC1 suggests that as PC1 increases, PC2 and PC3 decrease, or vice versa. Wait, the user says the clusters split along PC3 and PC2 while correlating negatively along PC1. Maybe the clusters are separated in the PC2-PC3 space, and PC1 is inversely related. So, for example, high PC1 values might correspond to lower PC2 and PC3, leading to different cluster regions.

What insights can we get? The clusters might represent different types of coffee profiles. For instance:

Cluster 1: High PC1 (strong aroma, flavor, etc.), low PC2 and PC3 (maybe less uniformity and sweetness). Cluster 2: Mid PC1, higher PC2 (more uniformity and sweetness). Cluster 3: High PC2 and PC3 (high uniformity and sweetness, but maybe lower PC1).

But since PC1 is inversely correlated with PC2 and PC3, clusters might be grouped based on their balance between overall quality (PC1) and specific attributes like uniformity and sweetness (PC2 and PC3). For example, Cluster 1 might be coffees with high overall quality but less uniformity or sweetness, while Cluster 3 has lower overall quality but higher uniformity and sweetness. Alternatively, the clusters could represent different market segments or processing methods, given that the original dataset has features like Processing.Method and Country.of.Origin.

Also, considering the original dataset's sample, the first sample has a high Cupper.Points and Total.Cup.Points, which might relate to PC1. The second sample has a higher altitude, which might influence the sensory attributes. However, the altitude is not part of the PCA features, so it's not directly captured here. The clusters might correspond to geographical regions (like Panama vs. Guatemala) or processing methods (natural vs. washed), which could affect the sensory attributes captured by the PCs.

Another angle: the negative correlation along PC1 suggests that as the overall quality (PC1) increases, the specific attributes (PC2 and PC3) might decrease. This could indicate a trade-off between general quality and specific characteristics. For example, coffees with high aroma and flavor (PC1) might have less uniformity or sweetness (PC2 and PC3), or vice versa. This could be due to factors like processing methods, which affect both the overall quality and the specific attributes.

The clusters could also represent different consumer preferences. For instance, Cluster 1 might appeal

to those who prefer strong, intense flavors, while Cluster 3 might be preferred for its smoothness and sweetness. The DBSCAN clustering would group coffees that are close in the PC2-PC3 space, which might correspond to similar sensory profiles.

Additionally, the presence of three clusters might indicate that the data has distinct groups based on these dimensions. The user should check if these clusters align with known categories in the dataset, like Country of Origin or Processing Method. For example, the sample data shows that Panama's Gesha has high scores, which might cluster in a certain PC space. Similarly, the higher altitude in some samples (like the third sample with 1600 meters) might influence the PC values, leading to different cluster formations.

In summary, the insights could be about identifying distinct coffee profiles based on a combination of sensory attributes, understanding trade-offs between different qualities, and relating clusters to factors like origin or processing methods. The user should validate these clusters against the original dataset's categorical variables to see if there are any patterns.

—THINKING STOP—

From the PCA and DBSCAN analysis, here are key insights and potential interpretations: 1. Dimensionality Reduction and Feature Prioritization

PC1 (highest loadings on Aroma, Flavor, Acidity, Body, Balance, Aftertaste) represents a core sensory quality of the coffee. It likely captures the intensity and balance of the coffee's flavor profile. High PC1 values may indicate coffees with strong, well-balanced, and aromatic characteristics (e.g., Gesha or high-altitude beans). PC2 (high loadings on Uniformity, Sweetness) and PC3 (similar loadings) reflect specific attributes like smoothness, consistency, and sweetness. These might indicate processing methods (e.g., washed vs. natural) or harvest quality. For instance, washed coffees (common in Guatemala) might cluster in PC2/PC3 due to their clean cup and sweetness, while natural-processed beans (like Panama Gesha) might have distinct PC2/PC3 profiles.

2. Clustering Interpretation (DBSCAN on PC2/PC3)

Cluster separation along PC2 and PC3 suggests distinct profiles in sweetness, uniformity, and consistency. For example: Cluster 1: High PC2/PC3 (sweet, uniform) → May correspond to washed coffees (e.g., the second sample from Guatemala) or high-quality, clean-cup beans. Cluster 2: Mid PC2/PC3 → Could represent moderate quality beans with balanced attributes. Cluster 3: Low PC2/PC3 (less sweet, less uniform) → Might include natural-processed beans (e.g., the first sample) or lower-altitude beans (e.g., the fourth sample with 3280m altitude). The negative correlation along PC1 implies a trade-off between overall quality (PC1) and specific attributes (PC2/PC3). For example: High PC1 (strong aroma/flavor) might correlate with lower sweetness/uniformity (e.g., the first sample with 81.83% moisture and 0.11% defects). Conversely, higher sweetness/uniformity (PC2/PC3) could come at the cost of slightly reduced overall intensity (PC1).

3. Practical Implications

Market Segmentation: Clusters could represent target consumer preferences: Cluster 1 (sweet, uniform): Appeals to customers seeking smooth, consistent profiles (e.g., espresso or filter brews). Cluster 3 (high PC1): Targets connoisseurs seeking intense, complex flavor (e.g., Gesha or anaerobic fermentation beans). Origin/Processing Insights: The third sample (Panama Gesha, 1600m altitude) has high PC1 and PC3, suggesting high-quality, sweet, and uniform beans. This aligns with the washed processing method and high-altitude (correlated with higher acidity/body in the dataset). The fourth sample (Guatemala, 3280m altitude) has high PC2/PC3 but lower PC1, possibly due to natural processing or lower moisture (75.58

4. Validation Opportunities

Correlate Clusters with Categorical Variables: Check if clusters align with Country of Origin (e.g., Panama vs. Guatemala) or Processing Method (washed vs. natural). For example: The sample with 1310.64m altitude (PC2/PC3) might cluster with Panama Gesha, while the 3280m altitude sample (PC2/PC3) might cluster with Guatemala Bourbon. Moisture and Defects: The moisture and Category One Defects (e.g., 0.11% in the first sample) could influence PC1/PC2/PC3. High moisture might reduce sweetness/uniformity (PC2/PC3), while defects could lower overall quality (PC1).

5. Actionable Takeaways

Optimize Processing Methods: If Cluster 3 (high PC1) is premium, prioritize washed processing for intensity. For sweetness, focus on natural-processed beans (Cluster 1). Altitude and Quality Trade-offs: Higher altitudes (e.g., 1600m) may enhance PC1 (intensity) but require balancing with PC2/PC3 (sweetness/uniformity) to avoid over-extraction or acidity. Target Markets: Use cluster profiles to segment marketing efforts—e.g., promote Cluster 1 as "smooth and sweet" for mainstream consumers, and Cluster 3 as "complex and intense" for specialty markets.

By aligning these insights with the dataset's categorical variables (e.g., Country, Processing Method, Altitude), you can further validate and refine the clustering's practical relevance.