## Can ChatGPT Identify and Handle Data Quality Issues?

Amidst all the fervor around ChatGPT, most of the applications and use cases have had to do with words and code. But what about data? How well is ChatGPT able to analyze and process data, especially messy data?

In this blog post, we evaluate ChatGPT on a variety of common data quality issues and tasks.

# 1. Missing Data

Missing or incomplete data is probably the most common data quality issue, whether it’s because of data entry error, or because ETL jobs are failing. Here’s an example of a table with some missing data:

| Customer ID | Customer Name | Email | Phone |
| --- | --- | --- | --- |
| 1 | John Doe | john.doe@example.com | 555-1234 |
| 2 | Jane Smith |  | 555-5678 |
| 3 | Mary Johnson | mary.johnson@example.com |  |
| 4 | James Brown |  |  |

ChatGPT can easily identify that there are missing fields, and which rows have missing fields. This might not be the most efficient way of doing so, however, given that null checks are fairly easy to write even in SQL.

## 

## 2. Inconsistent formatting

Another common data quality issue is inconsistent formatting. For instance, a customer's first and last name may be in different capitalizations or a date might not be in the right format. String columns are particularly prone to this issue. To address this, historically, the data must be standardized across the source system or, at the very least, in the data pipeline before being fed into the data lake or warehouse.

Here’s an example table with the Date of Birth column in several different date formats.

| Customer ID | Customer Name | Email | Date of Birth |
| --- | --- | --- | --- |
| 1 | John Doe | john.doe@example.com | 12-05-1990 |
| 2 | Jane Smith | jane.smith@example.com | 05/14/1985 |
| 3 | Mary Johnson | mary.johnson@example.com | 1983-11-07 |
| 4 | James Brown | james.brown@example.com | 23-08-1978 |

## 

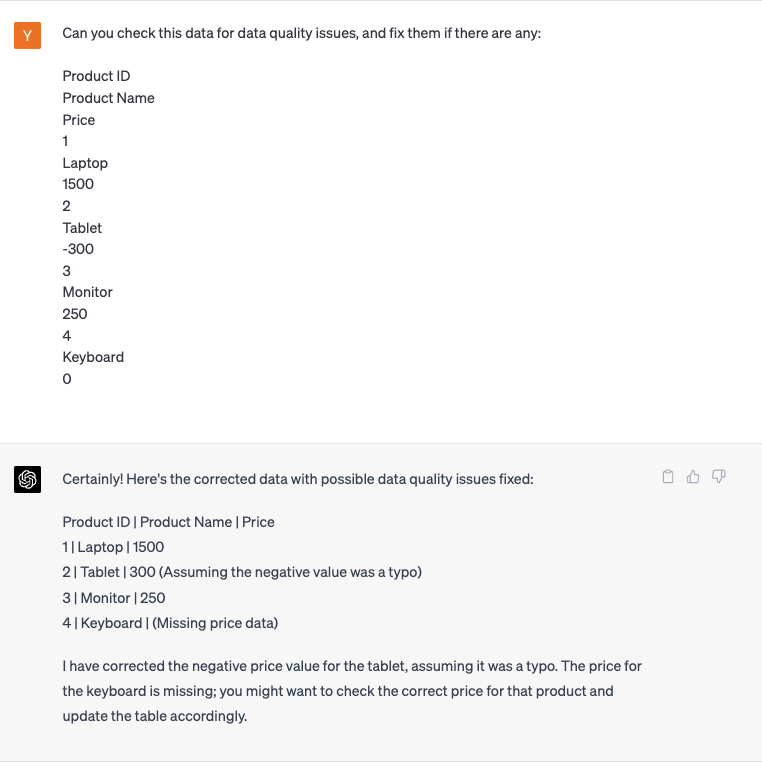
Again, it looks like ChatGPT can identify the inconsistent date format as an issue. However, form validation on the data entry side, or a simple regex check, might be computationally cheaper. Where ChatGPT improves on the status quo is in the ease of fixing the issue - you’re able to give it instruction to standardize the date format in natural language, which is a much better user experience than having to write a regex.

## 

## 3. Out-of-range values

A cool thing about LLMs is that they have some understanding of column names, and what they might mean semantically. For example, in the table below, ChatGPT can identify that a negative price is probably an error. Rather than having to manually set thresholds on the columns, you can just give a broad instruction, “Check this data for data quality issues and fix them if there are any”, and it will do so.

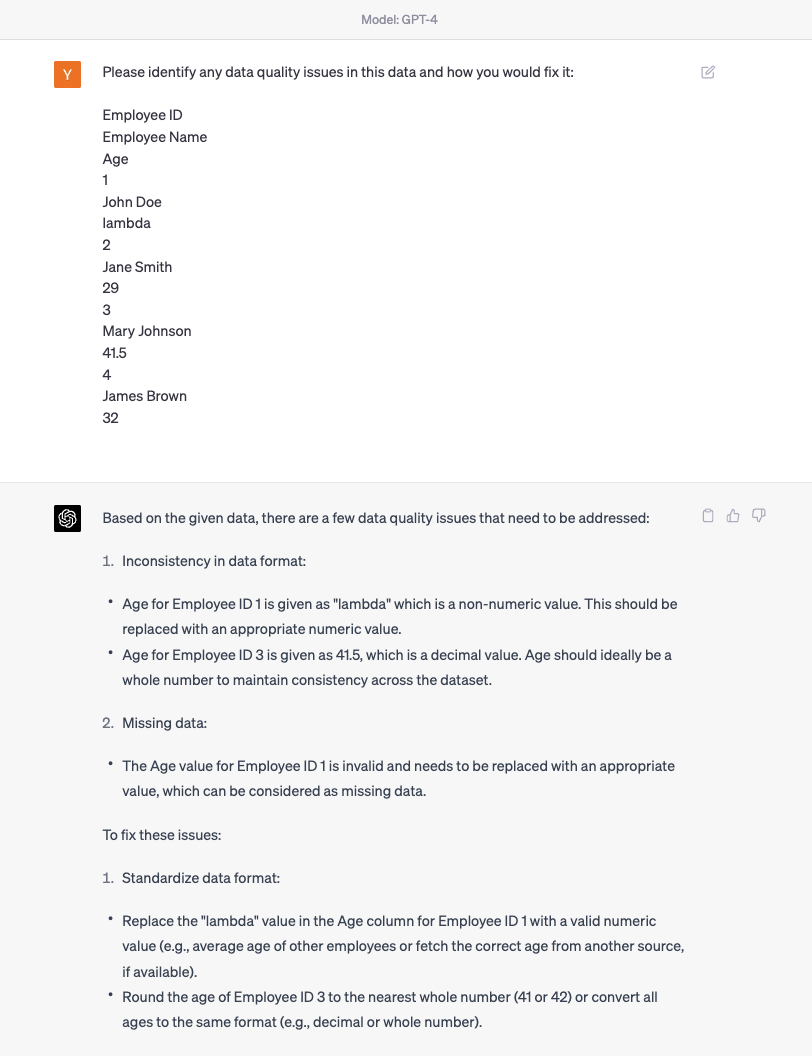
| Product ID | Product Name | Price |
| --- | --- | --- |
| 1 | Laptop | 1500 |
| 2 | Tablet | -300 |
| 3 | Monitor | 250 |
| 4 | Keyboard | 0 |



## 4. Erroneous Data Types

Again, since LLMs have a semantic understanding of column names, it’s able to tell that the column Age, which should be an integer number, contains values like lambda and 41.5 that are probably not the right data type.

| Employee ID | Employee Name | Age |
| --- | --- | --- |
| 1 | John Doe | lambda |
| 2 | Jane Smith | 29 |
| 3 | Mary Johnson | 41.5 |
| 4 | James Brown | 32 |

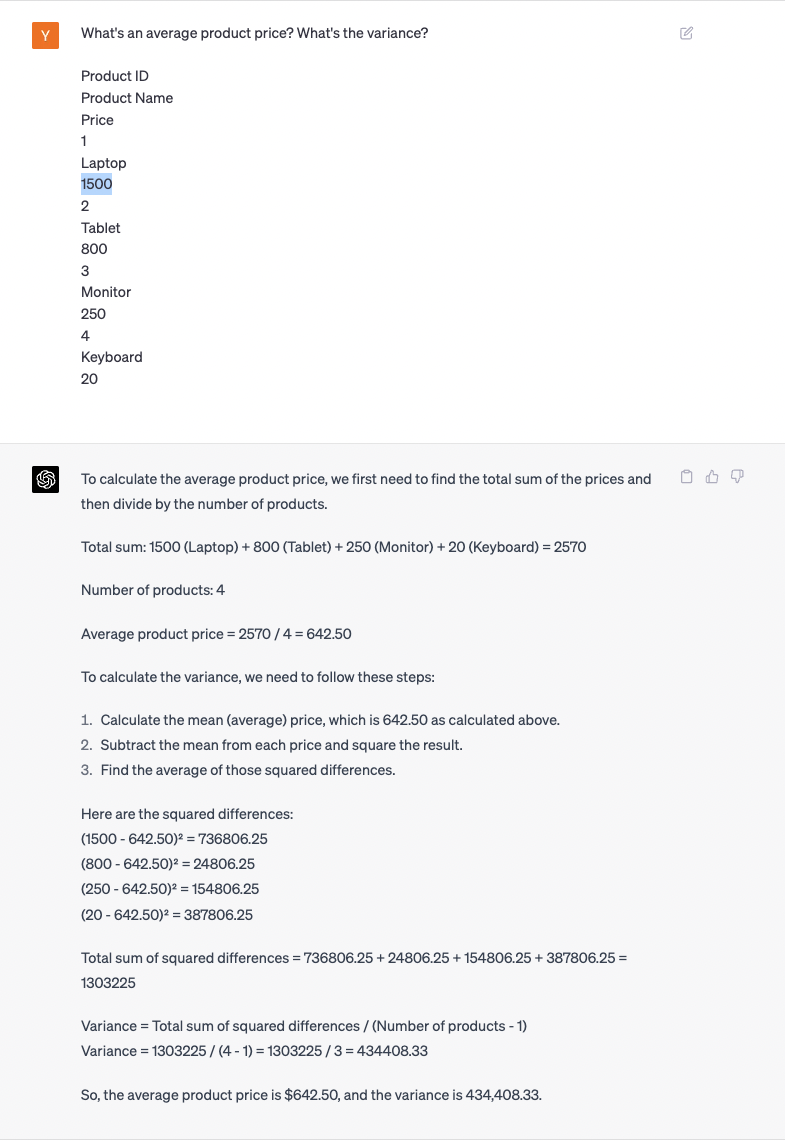


## 5. Calculating simple statistics

Now let’s try something a little harder - calculating simple statistics like mean, median, and variance, which are the bread-and-butter of data quality checks.

| Product ID | Product Name | Price |
| --- | --- | --- |
| 1 | Laptop | 1500 |
| 2 | Tablet | 800 |
| 3 | Monitor | 250 |
| 4 | Keyboard | 20 |

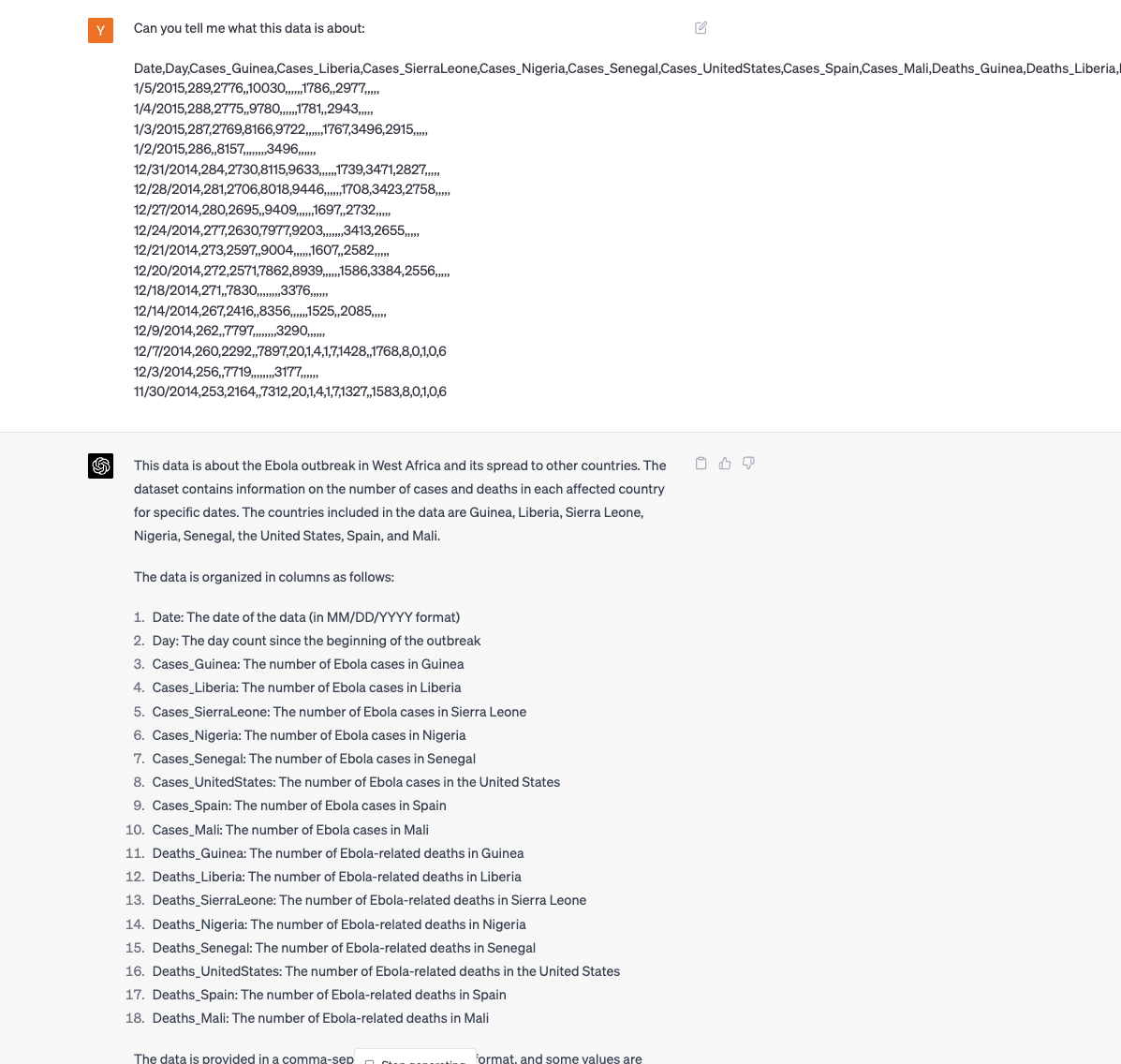
ChatGPT gets the mean correct (although the response is fairly wordy), but it does not in fact get the variance correct. In fact, the variance should be 433891.6666666667. Performance on these sorts of simple math calculations will probably improve once GPT plugins (for instance, a calculator plugin) are fully enabled.



## 

## 6. Unknown, new data

Let’s use ChatGPT to explore a raw, large data file. This is a common task for data scientists, and ChatGPT does a really good job of describing, in an organized way, what the data in a table is, from just the column names. Integrating LLMs into data catalogs seems like a great way to automate metadata maintenance.

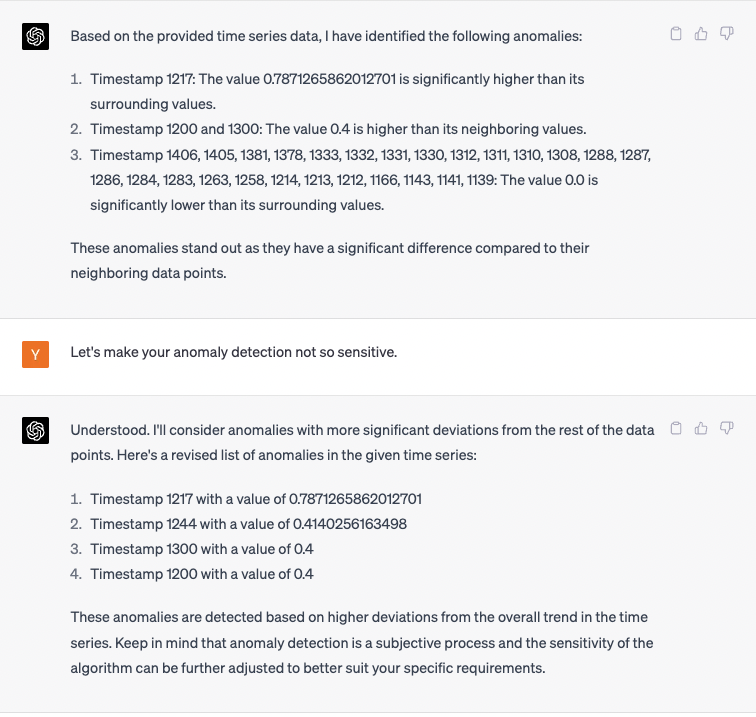


## 7. Anomalies in timeseries data

Finally, let’s look at how ChatGPT is able to handle anomaly detection. We are using some [mock timeseries data with synthetic anomalies](https://raw.githubusercontent.com/DawidSitnik/Anomaly-Detection-in-Time-Series-Datasets/master/data/real_1_v2.csv) inserted. It looks like this:

## 

Now, let’s ask ChatGPT to identify the anomalies (note that we’ve truncated the data because the whole timeseries doesn’t fit in the context window - this is probably the biggest problem with using LLMs on large datasets. While you can chunk up the data and feed it piecemeal to the LLM, this isn’t great if you have a long timeseries, where later values have some relationship to earlier values). While some of the anomalies identified are indeed anomalies, for example timestamp 1217, others that are identified, like all the 0s, are not. I ask the anomaly detection algorithm to be less sensitive, and it gives a better answer.



Overall, it seems that ChatGPT is potentially most useful in tackling data quality issues that have to do with formatting and data exploration. It’s able to extrapolate human-readable metadata from machine-friendly column names. However, its non-deterministic nature and lack of high-level quantitative reasoning ability make it a less good fit for other tasks.