This report is based on a 10,000 records based cytotoxic dataset where I had to perform various operations to find meaningful insights from the data.

# 1. Data preprocessing

In this stage I first removed unnecessary columns from the dataset from which I will not be able to get valuable information like ids,names, dates etc.

Then in the modified dataset I checked all the null and duplicate values by performing

# df.isnull().sum()

# df.duplicated().sum()

From my understanding chid and casn and dsstox are representing some unique id of the chemicals used,global\_mad and cytotox\_mad represent the median average data and cytotoxin\_median\_um and cytotoxin\_log\_median um represent the values in micrometer to measure toxicity and the last column is for reference to the date created

Here I saw that the cytotoxin\_median raw and global\_mad had null values around 6500

Then I wanted to identify those data that had hit more than 0 test samples.

Now I did the df.describe() and found out the following insights:

So I created a filter through

#df[df["nhit"]>0]

After this step I got a modified dataset with 4284 rows and no null or duplicated values.

Now I checked the number of unique chemicals in the dataset and found out that in all cases of nhit>0 all the chemicals were unique.

For this I used the operation

# df[df["nhit"]>0]["chnm"].nunique()

The output came as 4284 which matches the original number of rows in the dataset.

So now I move to the trend analysis part.

I also found out the extreme values (outliers) and found 20 rows worth outliers.

# 2. Trend Analysis

Here we analyze the trends of selective key columns in the data like cytotox\_median\_um,cytotox\_lower\_bound\_um and cytotox\_median\_log

We capture how cytotoxicity values are distributed without the extreme distortions from very rare high/low values.

After detecting and removing extreme cytotoxicity outliers, histograms revealed that most compounds cluster around moderate toxicity values (100–500 µM). The log-scale distribution appears approximately normal, indicating consistent experimental behavior. Lower-bound distributions align closely with median values, suggesting tight confidence intervals and robust measurement quality. This analysis highlights that strong cytotoxins are statistically uncommon, while the dataset mainly represents weakly toxic chemicals.

Now we see the most and least toxic chemicals by

# top10 = data.sort\_values('cytotox\_median\_um').head(10)

# least10 = data.sort\_values('cytotox\_median\_um').tail(10)

Now we explore relation between key attributes nhit,ntested and cytotoxin\_median\_um(micrometer) as evident from the graph.We observe that majority of the data is for cytotoxic materials as highlighted in purple but a few are less toxic chemicals highlighted as yellow.

# 3. Visualization

For visualization purposes, I have plotted the correlation heatmap and the barplot of the top and least 10 toxins and created a visual to find out which toxin had more and which toxin had the least toxic materials in it.

From the correlation heatmap we found out more about the key features and their proportionality relationship with each other.

# Summary

So summarizing everything we find out:

The project provides a clear statistical and visual understanding of the chemical dataset, highlighting the high redundancy of the cytotoxicity metrics, the inverse relationship between chemical potency and the number of recorded hits, and the specific identity of the most toxic compounds.