Python Analysis of Data from 2007 BMC Systems Biology article “Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes” by Bushel et al

Authors: Eric Mazlumyan, Isaac Brooks-Church

Summary

The excess data from Bushel et al (1) was analyzed and plotted against time and dose grouping to observe any significance in trends and determine the relative recovery of common liver damage markers. It was shown that, as Bushel et al mentioned, ALT levels underwent the most drastic rate of change when comparing low and high-dose rats. Our further analysis revealed other serum levels that showed similar rates of change and could be clinically significant. The levels of BUN, AST, TBA, and SDH all showed similar trends of a significant spike following the administration of high doses of acetaminophen. The expression of different significant genes and levels of liver condition severity were plotted and analyzed using PCA and k-means clustering, respectively, to uncover trends concerning the variables gene expression, liver condition severity, acetaminophen dosage, and treatment duration. Findings include dual clustering was appropriate for describing liver condition severity patterns.

Introduction

Principal component analysis (PCA) is a commonly used technique used to reduce the dimensions of very large experimental datasets. Very large datasets can be especially difficult to extract information from due to their size, which makes the dimensionality-reducing capacity of PCA very useful (2,3,4). PCA effectively filters out the least relevant data in a given experimental dataset by selecting for the data with the most variation in the dataset, thereby allowing the experimental data to be much more interpretable. PCA operates by constructing vectors called principal components (PCs) that correspond to data variables with the most variation and using the Euclidean space between PCs to reproject the most important experimental data in regards to the given experiment. Consequently, using PCA elucidates relationships among variables very effectively.

In their experiment, Bushel et al investigated the relationships between rat gene expression, acetaminophen treatment duration, acetaminophen dosage, liver condition severity, and rat serum levels. Data collection by Bushel et al resulted in obtaining a very large dataset. This is problematic because very large datasets make extracting biologically relevant information highly challenging. Using PCA on their experimental data is a worthwhile endeavor because it

would add clarity to their findings by allowing for the relationships between specific variables to be shown more clearly and resolving the very large dataset size issue.

The experiment that involved conducting PCA and k-means clustering on the liver condition data is significant because it allowed for clearly showing the relationship between liver condition severity and number of acetaminophen treatment hours. K-means clustering is a useful technique because it allows for elucidating the similarities between different data points in a given experimental dataset (5). K-means clustering was a significant step to make because it elucidated similarities between the effects of different numbers of hours of acetaminophen treatment on liver condition severity.

The clusters in k-means clustering are groups of data points that show similarity whose centers can be determined. The number of clusters is determined using the elbow method. The elbow method considers the relationship between inertia, which represents the degree to which data points are far from their closest cluster center, and number of clusters. The greater the number of clusters, the smaller the value for inertia since data points have a greater intrinsic ease to be closer to a cluster center with more clusters present. Incrementally increasing the number of clusters reveals the most appropriate number of clusters for a given dataset because the most appropriate number of clusters for a given dataset corresponds to the most drastic decrease in the extent to which inertia decreases as the number of clusters increases. In other words, the "elbow" in the relationship between inertia and number of clusters indicates the most appropriate number of clusters.

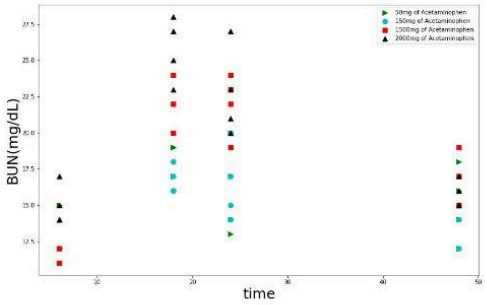
The over-the-counter pain reliever acetaminophen is recommended to not exceed 15 mg/kg by IV administration every 6 hours. The LD50 was discovered to be greater than 2000 mg/kg of body weight. This shows how the typical daily recommendations are quite conservative. Death is however quite different than liver necrosis. This experiment assays the different levels of gene expression and serum levels in rats after severe doses of acetaminophen were administered. This allows the researcher to observe how much acetaminophen (below the LD50) can be administered before liver necrosis and severe damage is observed. The data was obtained from rats, however, it can be cautiously extrapolated to humans as a basis for further experiments.

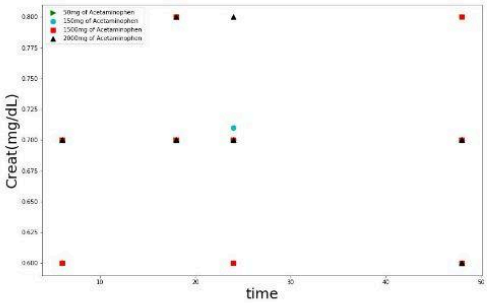
When analyzing the data presented by Bushel et al certain data sets were found to have been excluded for the sake of brevity or due to a lack of significant results. The main figure of their work, while very condensed and accessible, lacked a deeper analysis which we obtained through PCA. We also took all of the "junk data" and ran it through several comparisons to determine if any correlations were observable between serum levels, dosage, and time. This analysis provides a deeper look at physiological changes and gene expression after high doses of analgesic or antipyretic drugs are administered and give more insight that can be used to advise usage in humans.

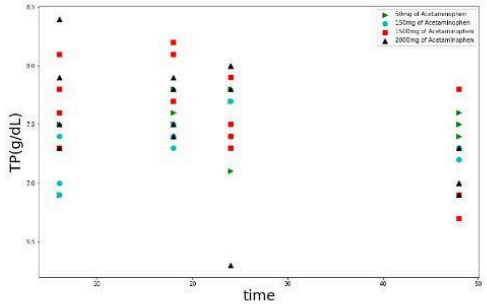
Analysis

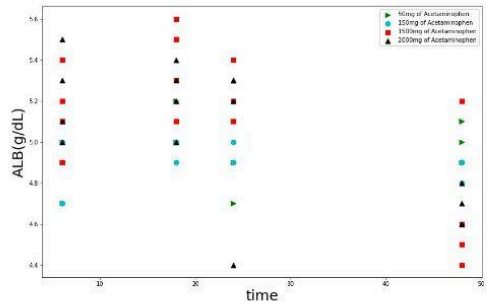
To further analyze the rat phenotypes, the rat data for their full serum measurements were extracted from the paper and plotted. To do this the raw data was downloaded and put into a spreadsheet software, which was followed by organizing the data and removing string symbols. Once the data was organized and imported properly it was plotted by time group and dosage group (Figure 1A). Due to the nature of the experiment, the same rat could not be used twice since the liver damage would confound later results. To further simplify the data the average serum levels for the four different rats at each time point and dosage group were averaged and plotted against the other rat averages per time and dose (Figure 1B).

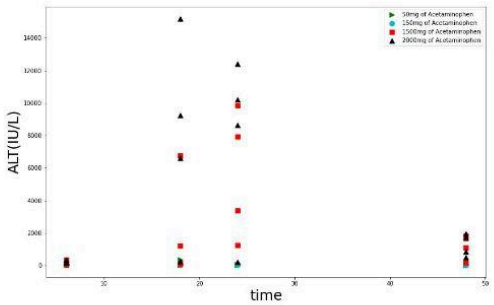
16 Rats Given Acetaminophen over time

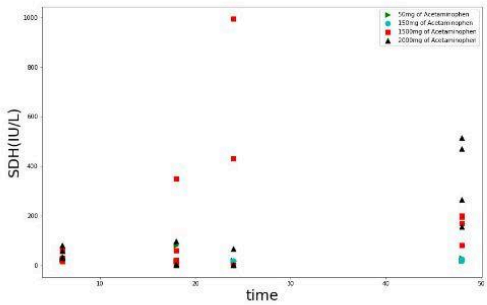


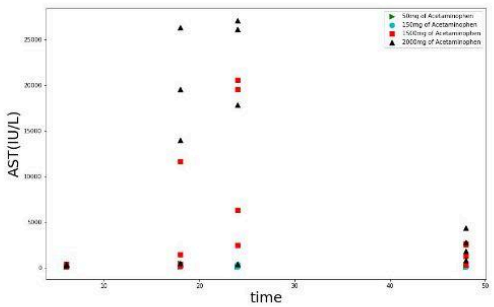


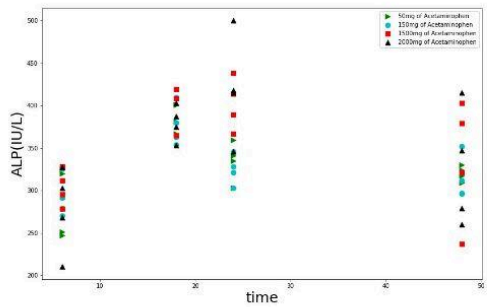


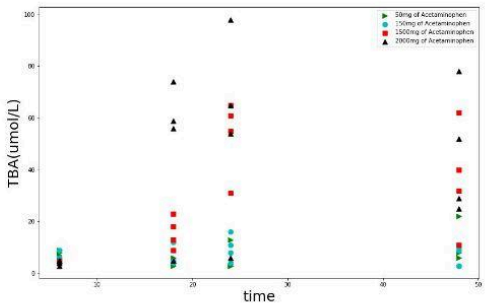












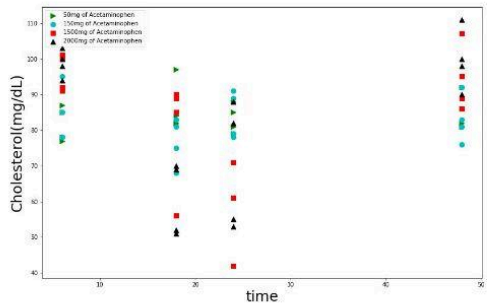
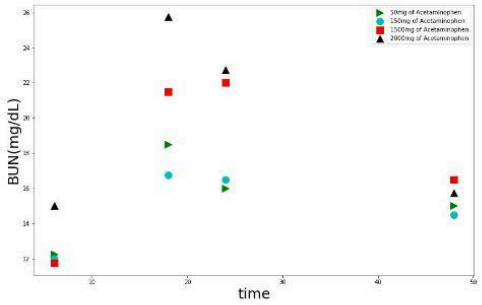
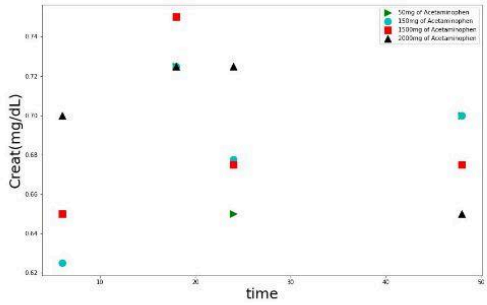


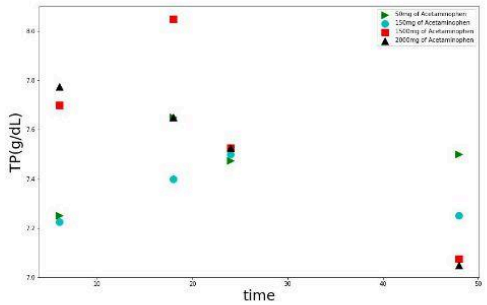
Figure 1A

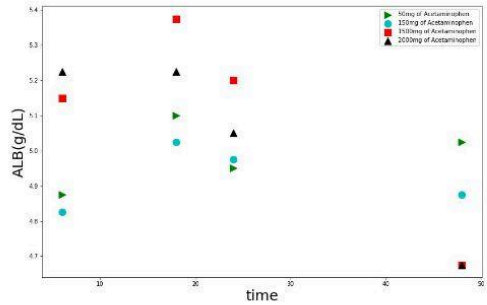
Serum levels of 64 different rats taken at specific doses and time points. The data used for this analysis was taken from the unused datasets in the original paper. Each marker is generally indicative of overall kidney health. The rats were observed at four different time intervals; the blue, yellow, green, and red data points indicate the rats given doses of 50, 150, 1500, and 2000 mg/kg of body weight.

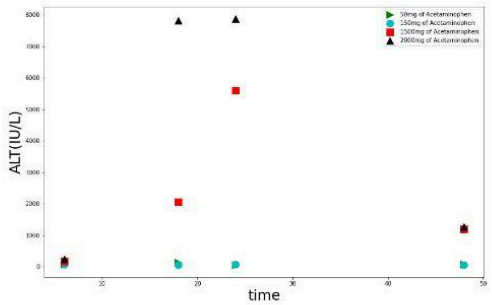
16 Rats Given Acetaminophen over time, average values at time

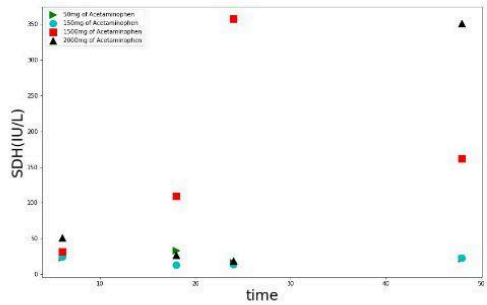


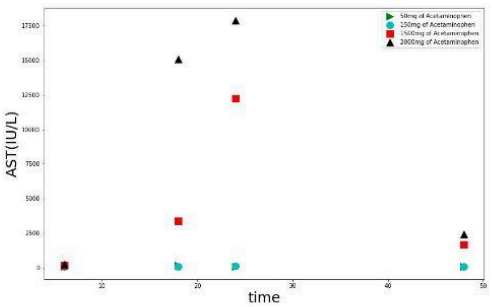


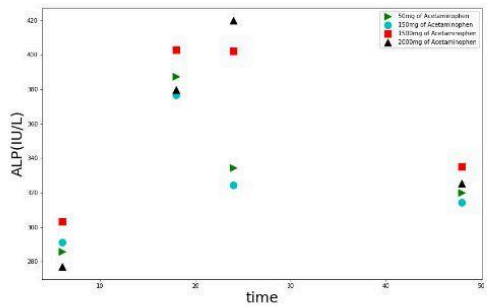


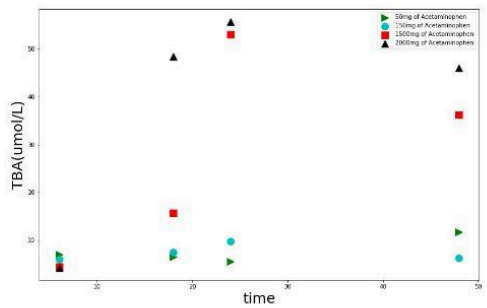












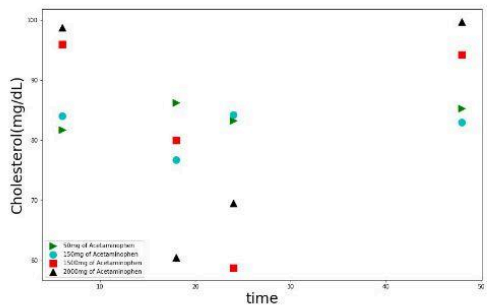


Figure 1B

Serum levels of 64 different rats taken at specific doses and time points. To increase clarity of Figure 1A, the average serum levels of the rats of each respective time group and dose group were taken and plotted. Each marker is generally indicative of overall kidney health. The rats were observed at four different time intervals; the blue, yellow, green, and red data points indicate the rats given doses of 50, 150, 1500, and 2000 mg/kg of body weight.

The blood urea nitrogen levels for all dose groups rose as the dosage increased with the level beginning to go down after 12 hours. The creatine levels for the rats showed no significant trends. The total protein levels seemed to initially peak for the rats given 1500 mg of acetaminophen or more. However, after 48 hours the rats given lower doses scored higher TP levels. The levels of albumin followed the same trend as those from the total protein measurements. The alanine transaminase results show a much more distinct disparity between rats given more than 1500 mg per kg and those given lower doses. The spike of ALT levels stayed similar until after 6 hours when a significant spike was observed in higher-dosed rats. The levels recovered slowly after 12 hours however notably the ALT levels between different dose groups were still quite distinct at 48 hours. Rat succinate dehydrogenase, aspartate aminotransferase, and total bile acid assays results followed a similar trend to ALT. The alkaline phosphatase levels also indicated a response to higher doses of acetaminophen however it was not as drastic as the previous results. The cholesterol levels of the rats also seemed to respond to higher doses, however, the relationship was inverse with higher doses correlating with lower cholesterol levels.

Of the original 3100 genes 82 genes were selected as being significantly impactful in predicting rat liver necrosis by performing a chi-square test. The raw data unfortunately was listed using the micro-array control type rather than identifiable gene labels. In order to decode the data, the product page for Agilent Technology, which was credited in the experiment, was searched. The specific microarray (G4130A) was not found on the official website. Doing a search on the Gene Expression Omnibus, however, did give a key for the specific micro-array (6). Using this key the raw data was successfully analyzed and organized according to the significant genes Bushel et al had discovered. Data that Bushel et al. collected but did not use were involved in the independent experiments.

PCA was performed using the experimental gene expression data by mean-centering the data, extracting the eigenvalues and eigenvectors from the data, determining the reprojected data matrix, and plotting the reprojected data using the first and second PCs. The initial organization of the experimental dataset was conducive for setting the first PC to capture the variance of treatment duration. The second PC captured the variance of gene expression. Data points were labeled by acetaminophen dosage to depict its influence as a third variable. Figure 2A revealed that durations of 24 and 48 hours were associated with greater levels of gene expression.

A second PCA was conducted with an exchange for the variable corresponding to the first PC from treatment duration to acetaminophen dosage because it provided an additional way to observe the relationship between variables and capitalized on the usefulness of the PCs in analyzing reprojected data by using the role of the first PC in more than one way. The experimental dataset was rearranged as needed for the second PCA. Data points were labeled by treatment duration to depict its influence as a third variable. Figure 2B revealed that dosages of 50 and 150 mg/kg were associated with intermediate levels of gene expression.

PCA was performed again using the experimental liver condition data, with treatment duration and liver condition severity as the variables used for PC assembly. Data points were

labeled by acetaminophen dosage to depict its influence as a third variable. Figure 3A revealed that durations of 24 and 48 hours were associated with greater levels of liver condition severity.

The process of k-means clustering began by simplifying Figure 3A to only consider the variables of treatment duration and liver condition severity (Figure 3B) such that the information extracted from the cluster centers is highlighted. The elbow method was utilized by graphing inertia values against their corresponding number of clusters for the reprojected liver condition data. The “elbow” depicted by the plot (Figure 3C) was associated with a cluster number of 2. In Figure 3D, the two cluster centers were plotted along with the plotted data depicted in Figure 3B. One cluster center was located closer to data points corresponding to durations other than 6 hours, and one cluster center was located closer to data points corresponding to 6 duration hours.

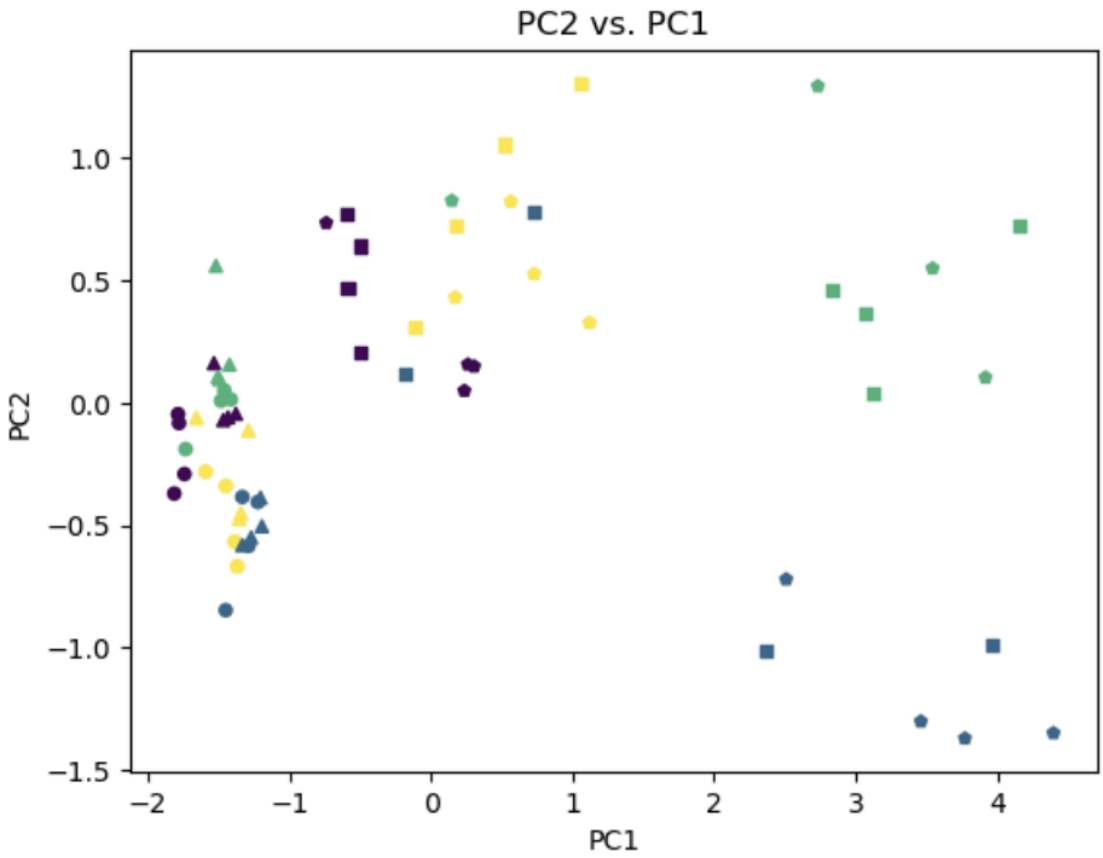


Figure 2A

PC capturing variation of treatment duration (PC1) versus PC capturing variation of gene expression (PC2). Purple, blue, green, and yellow represent 6, 18, 24, and 48 hours, respectively. Circles, triangles, squares, and pentagons represent acetaminophen dosages of 50, 150, 1500, and 2000 mg/kg, respectively.

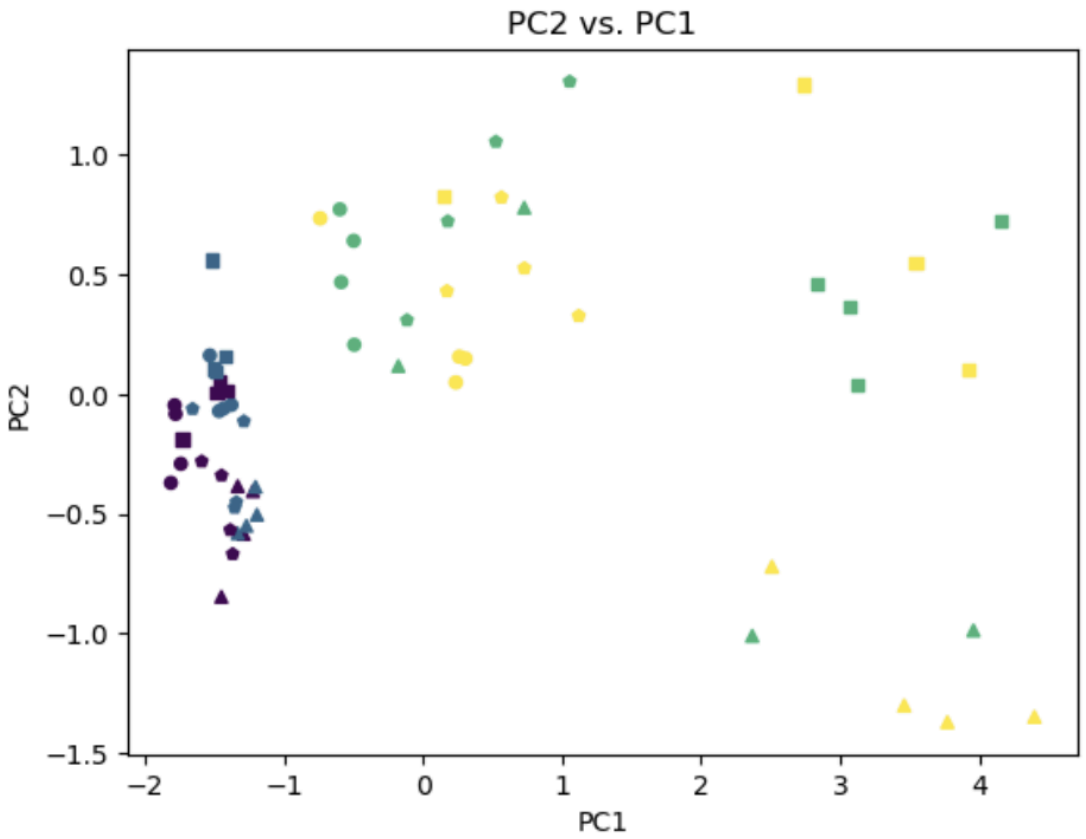


Figure 2B

PC capturing variation of acetaminophen dosage (PC1) versus PC capturing variation of gene expression (PC2). Purple, blue, green, and yellow represent 50, 150, 1500, and 2000 mg/kg, respectively. Circles, triangles, squares, and pentagons represent treatment duration hours of 6, 18, 24, and 48 hours, respectively.

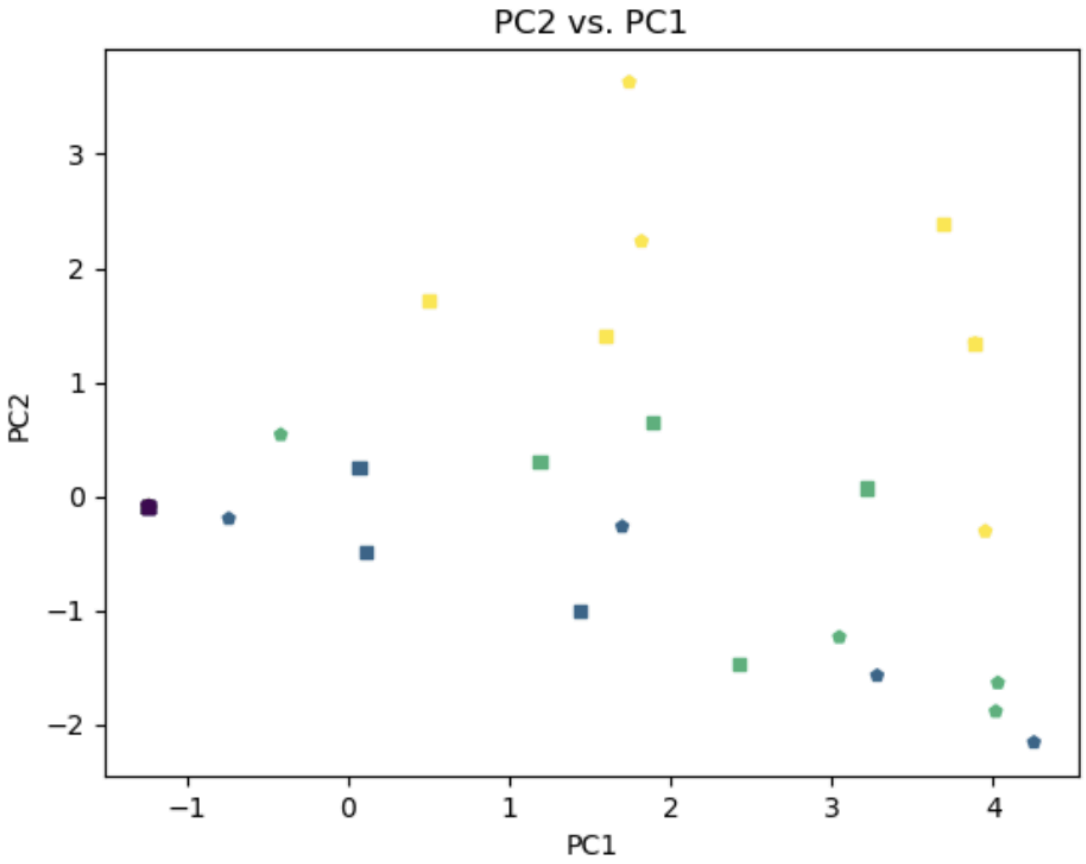


Figure 3A
PC capturing variation of treatment duration (PC1) versus PC capturing variation of liver condition severity (PC2). Purple, blue, green, and yellow represent 6, 18, 24, and 48 hours, respectively. Circles, triangles, squares, and pentagons represent acetaminophen dosages of 50, 150, 1500, and 2000 mg/kg, respectively.

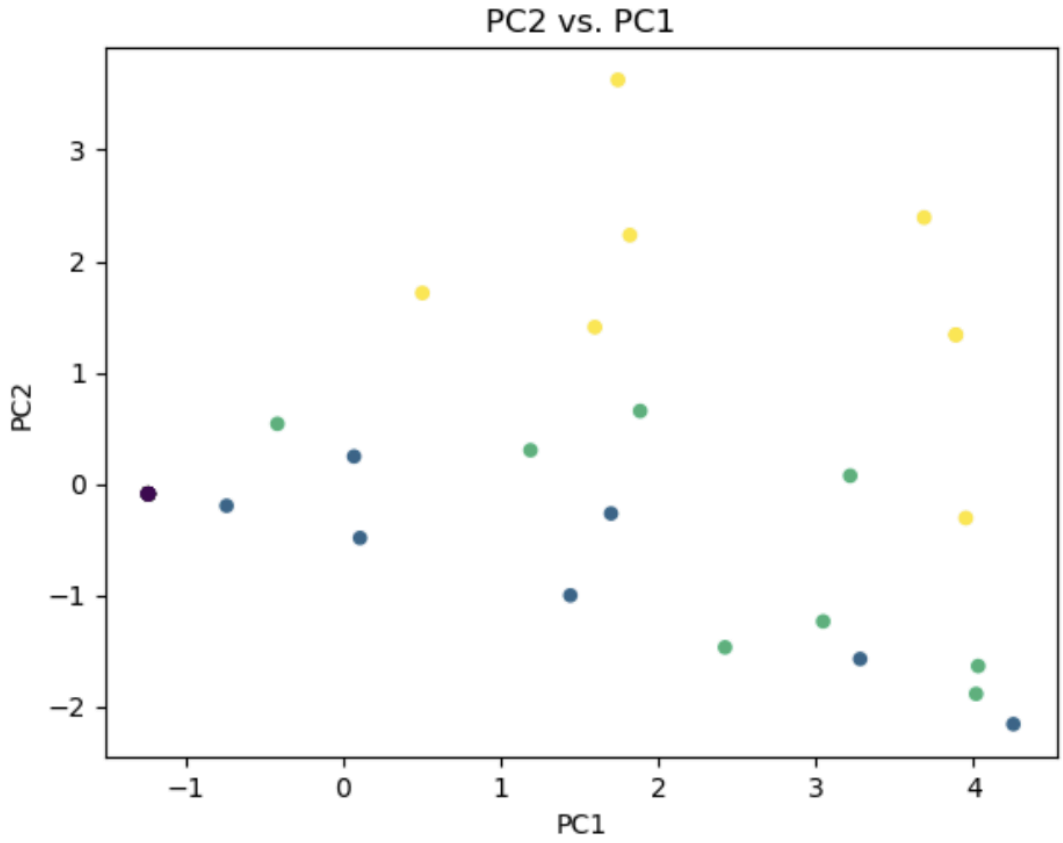


Figure 3B
Simplified version of Figure 3A.

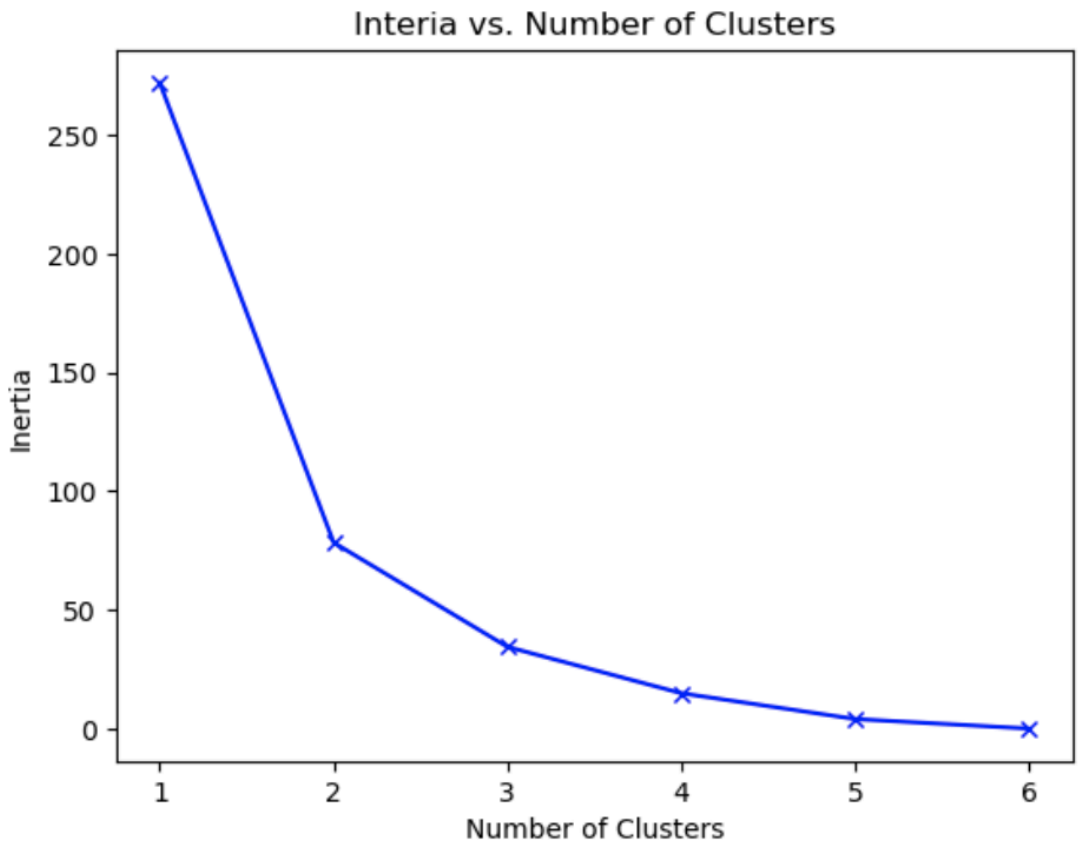


Figure 3C
Inertia versus number of clusters using the reprojected liver condition data. “Elbow” of plot is associated with cluster number of 2.

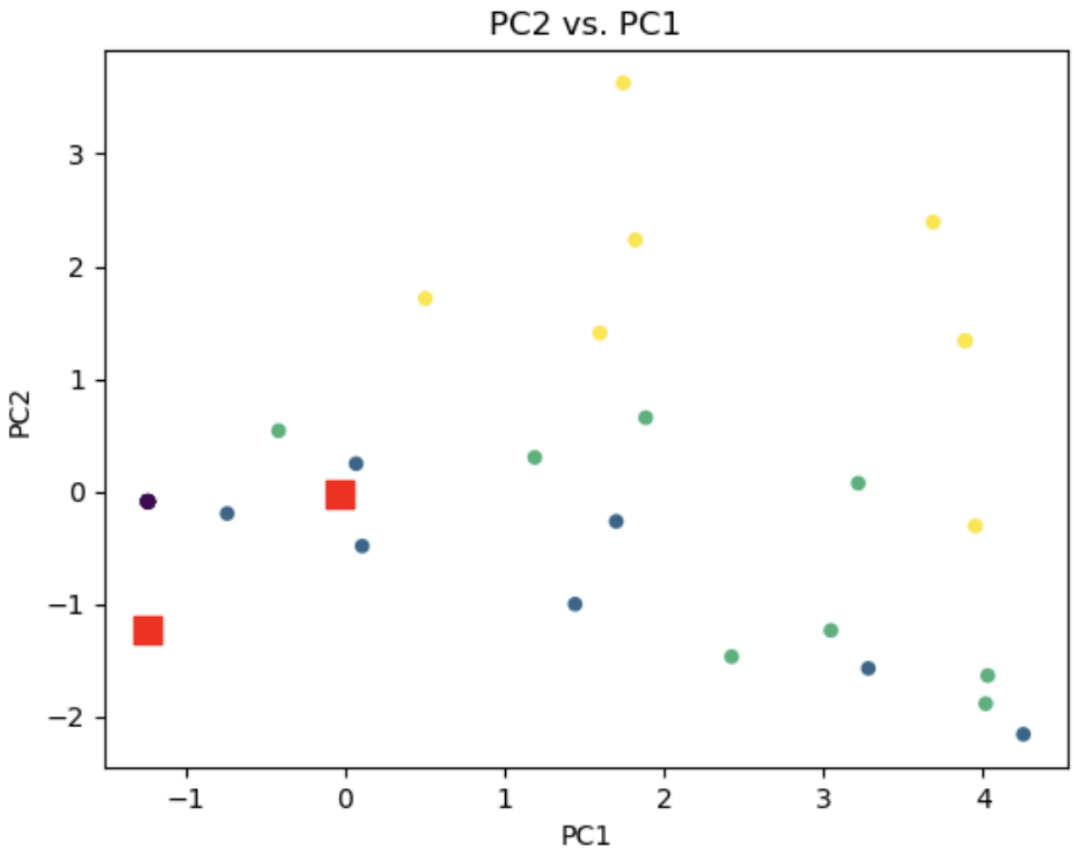


Figure 3D
Modified version of Figure 3B with cluster centers depicted as red squares.

Discussion

When looking at the serum levels for the rats a few standouts are apparent. The most dramatic and significant observation is that of the ALT levels. ALT is released into the bloodstream after inflammation of the liver occurs. Analgesic and antipyretic drugs, like acetaminophen, are known to be hepatotoxic at higher doses and this is borne out in the ALT assay. There seems to be an upper threshold of toxicity based on the data. Rats given 2000 mg/kg were seen to peak in their levels of ALT at the 12-hour measurement with not much of a change (based on averages) at the 18-hour mark. This is contrasted with rats given 1500 mg/kg which were observed to still be rising at the 18-hour mark. This indicates that at a certain point, the levels of liver toxicity (based on ALT released) may reach a limit. In this data, there were some outliers with some rats' ALT only elevating slightly in the 1500 mg/kg group. Noticeably the levels of liver inflammation did not seem to have fully recovered after 48 hours for all high-dose groups. The BUN, TBA, AST, and SDH levels for all groups also followed a similar trend to ALT. The only insignificant result was that of the creatine levels. Since creatine levels can be altered by diet, muscle mass, as well as liver stress it explains why this data set was so mixed.

The results from Figure 2A indicated that, as treatment duration increased, gene expression underwent a decrease that was followed by an increase. This suggests that strictly relatively large increases in treatment duration causes increases in gene expression. The results from Figure 2B indicated that, as acetaminophen dosage increased, the range of the level of gene expression conducted increased. This suggests that, as acetaminophen dosage increases, the expressions of specific genes are upregulated more and the expressions of other specific genes are downgraded more. The results from Figure 3A strongly indicated that, as treatment duration increased, liver condition severity increased. This strongly suggests that there is a direct relationship in which increases in treatment duration cause increases in liver condition severity. The results from Figure 3C indicated that the data can be grouped into two groups based on degree of similarity. This suggests that the effect of treatment duration on liver condition severity can be characterized by two different patterns. The cluster centers depicted in Figure 3D indicated that the data points corresponding to 6 duration hours are relatively similar to each other, the data points corresponding to durations other than 6 hours are relatively similar to each other, and the two groups of data points are relatively different. This suggests that 6 duration hours has a relatively different effect on liver condition severity compared to the other durations.

From this information, it can be gleaned that treatment with acetaminophen for longer than 6 hours can have drastic consequences on liver health and that treatment with acetaminophen is a time-sensitive matter. Additional independent experiment steps that could be taken include conducting PCA with PC1 capturing variation of acetaminophen dosage in place of treatment duration in order to provide an additional way to observe the relationship between acetaminophen dosage and liver condition severity. Future directions could include introducing a new variable, like diet, and observing the effects of diet on liver condition severity after conducting PCA.

As an easily accessible OTC medication, acetaminophen and its effects are important to fully understand. The data presented in this paper and the analysis performed here demonstrate the potential liver necrosis that can occur if it is abused.

While these results are significant this experiment could have utilized control data taken at time zero. This would give a good comparison and alleviate some errors that may arise through pure data comparison without a baseline measurement. For example, when comparing the rat serum levels to the normal reference ranges noted in other scientific papers (7,8,9), the discrepancies could be quite extreme. This makes sense considering these rats were subjected to high doses of analgesic drugs, however, without a baseline it is difficult to say how much these levels changed when compared to the normal markers.

Works Cited

1. Bushel, P.R., Wolfinger, R.D. & Gibson, G. Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. *BMC Syst Biol* 1, 15 (2007). https://doi.org/10.1186/1752-0509-1-15
2. Ringnér, M. What is principal component analysis?. *Nat Biotechnol* 26, 303–304 (2008). https://doi.org/10.1038/nbt0308-303
3. Yao, F., Coquery, J. & Lê Cao, KA. Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets. *BMC Bioinformatics* 13, 24 (2012). https://doi.org/10.1186/1471-2105-13-24
4. Jolliffe, Ian T., and Jorge Cadima. "Principal component analysis: a review and recent developments." *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016): 20150202.
5. Education Ecosystem (LEDU). "Understanding K-means Clustering in Machine Learning." *Medium*, Towards Data Science. https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1. Accessed 23 Mar. 2023.
6. Agilent Technologies. *Agilent-011868 Rat Oligo Microarray G4130A (Feature Number version)*. 2004. *Gene Expression Omnibus (GEO)*, National Center for Biotechnology Information, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?view=data&acc=GPL890&id=20440&db=GeoDb_blob82
7. Thammitiyagodage, M.G., de Silva, N.R., Rathnayake, C. *et al.* Biochemical and histopathological changes in Wistar rats after consumption of boiled and un-boiled water from high and low disease prevalent areas for chronic kidney disease of unknown etiology (CKDu) in north Central Province (NCP) and its comparison with low disease prevalent Colombo, Sri Lanka. *BMC Nephrol* 21, 38 (2020). https://doi.org/10.1186/s12882-020-1693-3
8. Owu DU, Osim EE, Ebong PE. Serum liver enzymes profile of Wistar rats following chronic consumption of fresh or oxidized palm oil diets. Acta Trop. 1998 Mar;69(1):65-73. doi: 10.1016/s0001-706x(97)00115-0. PMID: 9588242.
9. Siques P, Brito J, Naveas N, Pulido R, De la Cruz JJ, Mamani M, León-Velarde F. Plasma and liver lipid profiles in rats exposed to chronic hypobaric hypoxia: changes in metabolic pathways. High Alt Med Biol. 2014 Sep;15(3):388-95. doi: 10.1089/ham.2013.1134. Epub 2014 Sep 3. PMID: 25185022; PMCID: PMC4175031.