**Analysis**

**Where did the data come from?**

It is a project developed in the Broad Institute of MIT and Harvard, the Laboratory for Innovation Science at Harvard (LISH), and also the NIH Common Funds Library of Integrated Network-Based Cellular Signatures (LINCS), offered this data reservoir with the purpose of developing drug advancement through enhancements to Mechanisms of Action (MoA) prediction algorithms. We have chosen this dataset from Kaggle which is an open-source community of data scientists and machine learning enthusiasts

**Why did you choose this data?**

In the past, scientists determined drugs from natural and essential products or were motivated by conventional remedies. Quite popular drugs, like paracetamol, known in the US as acetaminophen, were set into clinical practice decades before the biological mechanisms starting their pharmacological activities were understood. Now, with the advent of more robust technologies, drug development and detection have been adapted from the unexpected paths of the past to a wider targeted model that fostered a perception of the underlying biological mechanism of a disease or a virus. In this, experts seek to detect a protein target associated with disease and generate a molecule that will accentuate the protein target. In short, scientists designate a name or a tag remarked as Mechanism of Action or MoA

**What did you do with the data in the context of exploration?**

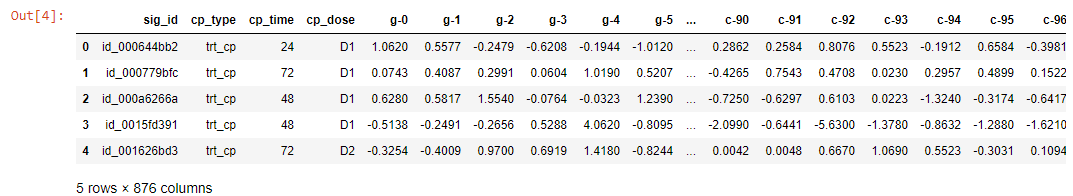
We have done some basic research on these mechanisms, genes, drugs, and cells. Also, worked on the Exploratory Data Analysis (EDA) and used Python to plot the visualizations, some basic calculations and analysis. We also checked for the null values, data anomalies, and duplicates.

**What did you find?**

**train\_features.csv**

Features like

* g – signifies the gene expression data
* c – signifies the cell viability data
* Number of rows in training set: 23814
* Number of columns in training set: 876

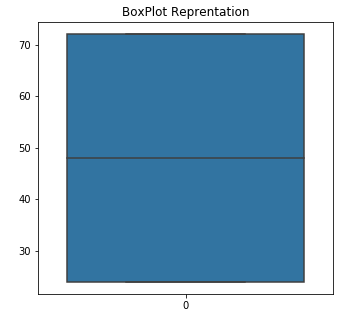
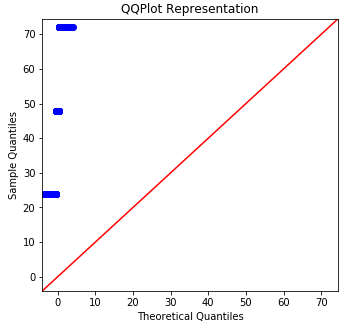


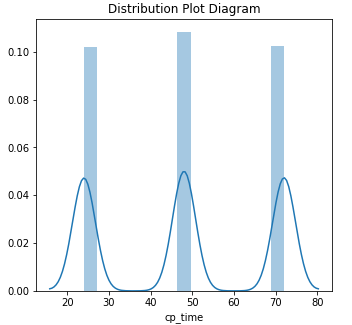
There are 3 categorical features mainly like cp\_type, cp\_time, and cp\_dose. Two of them are binary features and one amongst them has three unique values, that the cardinality among those features is extremely low. All of the distinct features have substantially identical patterns in training and public test set, which implies training and public test set are obtained from the same sample. Samples are probably stratified on these features while splitting training and public test set, and therefore the same distributions are foreseen within the private test set.

* **cp\_type** is that the first categorical feature within the dataset and it's a binary feature. It either means samples are treated with a compound (trt\_cp) or with a sway perturbation (ctl\_vehicle). Samples treated with control perturbations haven't any MoAs, thus all of their scored and non-scored target labels are zeros. However, all zero-labeled samples don't seem to be entirely treated with a sway perturbation, over 1/3 of the compound samples also are labeled as zeros.

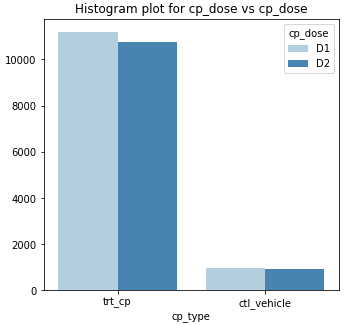


* **cp\_time** is that the 2nd categorical feature within the dataset and it has 3 unique values they are 24, 48, and 72 hours. It indicates the treatment durations of the samples. Sample counts of various cp\_time values are very uniform and closed to every other in numerous targets. Sample counts are either extremely near one another or 48 is somewhat more than the others.

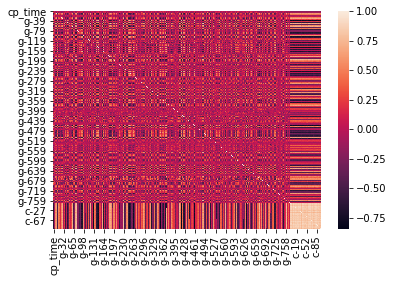
 

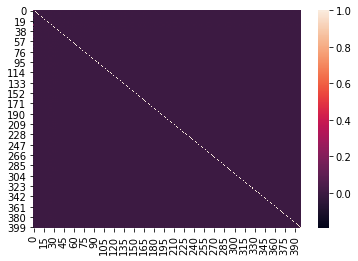


* **cp\_dose** is that the final categorical feature within the dataset and it's also a binary feature. It indicates whether the dose of the samples is either low (D1) or high (D2). Sample counts of various cp\_dose values are very consistent and shut to every other in numerous targets. Sample counts are even closer to every other compared to cp\_time and much of targets have equal sample counts for both doses. All targets have samples with two doses except erbb2 inhibitor and atp-sensitive potassium channel antagonist because those two targets are only classified in one sample.

These features are related to cell viability and gene expression.





**Why does that matter?**

**Cell Viability**

Cell viability is a measure of the proportion of live, healthy cells within a population. Cell Viability is an essay that is designed to discover the ability of organs, cells, or tissues to maintain or improve a state of endurance. Viability can be seen from the all or nothing situations of life and death by the use of a quantifiable index that varies between the integers of 0 and 1 or if more easily explained, the range of 0% and 100%.

Viability can be observed through the physical properties of cells, tissues, and organs. Some of these include mechanical activity, motility, such as with spermatozoa and granulocytes, the contraction of muscle tissue or cells, mitotic activity in cellular functions, and more. Viability assays provide a more precise basis for measurement of an organism's level of vitality.

Cell Viability assessment is based on PRISM technology. PRISM is a high-throughput screen for assessing cell viability in which cell lines that have each been labeled with a unique 24-nucleotide barcode are pooled and treated with the experimental condition, and surviving cells are “counted” through the classification of the related barcode. PRISM is an acronym for Profiling Relative Inhibition Simultaneously in Mixture.

Cell viability should range between integers 0 and 1. Here, we have values in the range of 10 and 6 because the data were z-scored and then normalized using a scheme called quantile normalization.

* High negative values = High number of dead cells
* High positive values = High number of living cells

There are 100 cell-viability features and they have c- prefix (c-0 to c-99). Each cell-viability feature represents the viability of one particular cell line, and all experiments are based on a set of similar cells. These are mostly cancer cells.

**Gene Expression**

It is the method by which information from a gene is employed within the synthesis of a functional gene product. These products are often proteins. In short, the mechanism of action of the 207 targets during this study will activate some genes, organic phenomenon will occur, and byproducts (proteins) are synthesized. There are 772 organic phenomenon features and that they have g- prefix (g-0 to g-771). Each organic phenomenon feature represents the expression of 1 particular gene, so there are 772 individual genes are being monitored during this assay. The mean organic phenomenon of the 772 genes shows some strong negative and positive organic phenomenon values. this could flow from to several factors:

* Some drugs upregulate and others downregulate some genes: For example, drug-1 could reduce gene-A expression level while drug-2 could elevate gene-B expression level.
* Some genes have high negative correlation, so gene-A has a high positive gene expression value which means gene-B will have a high negative gene expression value.

**What would your proposed next steps be?**

We determine the MoAs of a replacement drug and one approach is to treat a sample of human cells with the drug and so analyze the cellular responses with algorithms that seek for similarity to known patterns in large genomic databases, like libraries of gene expression or cell viability patterns of drugs with known MoAs.

**What business problem are you intending to solve using ML with the data?**

To use the training dataset to develop an algorithm that automatically labels each case in the test set as one or more MoA classes.