# t-SNE and K-Means Clustering on Drink Dataset

The task at hand is to find if there are natural clusters in the Drink dataset and if there are, to interpret these clusters to find appropriate labels for the dataset or subsets of the dataset.

## Data Preparation

In the ‘Drink’ dataset there are 11 variables. All of these are continuous numerical variables.

**Step 1** : To understand dataset variables(dimensions) and to find if there are any variables which are providing the same information, that can be dropped.

To do that, plotting correlation heatmap –

A screenshot of a video game

Description automatically generated

From the heat map, listing the variable pairs with higher positive/negative correlation:

**Density-sugar = 0.82**

**Density-alcohol = -0.78**

**Density-total sulfur dioxide = 0.53**

**Free sulfur dioxide-total sulfur dioxide = 0.62**

**NOTE: The area/business of concern for this dataset is a lot related to chemicals of which we have minimal knowledge.**

But we know that density of any liquid (assuming from the dataset title ‘Drink’) will increase if we put dense substances like residual sugar or total sulfur dioxide in it. So, it is safe to drop ‘density’ variable since it has high correlation with three variables and the cause for correlation is clearly due to correlated variables (sugar, total sulfur dioxide, alcohol)

Another high correlation is between total sulfur and free sulfur dioxide, which, due to uncertainty of the chemicals, are not being dropped.

**Step 2:** To find natural clusters in this unfamiliar-variables dataset, we need to form subsets. Subsets descriptions are explained in section below.

Also, to interpret the clusters and to assign appropriate labels, ‘converter’ functions have been defined throughout the code.

## t-SNE Implementation

t-SNE is an unsupervised machine learning, non-linear dimension reduction technique which is used to visualise large multi-dimensional datasets into two or three dimensions. The t-SNE algorithm converts Gaussian distribution to Cauchy distribution and by using CAUCHY distribution, reduce the tendency of crowding data points in low-dimensional representation.

**Reason for using t-SNE**

In the given dataset there no labels or target variable. In order to find appropriate labels, we need to find natural clusters or patterns in the dataset which have 11 dimensions. To find patterns, first we need to visualise the dataset. Visualisation of 11 dimensions is not possible and hence we need to use dimension reduction technique.

There are two options when it comes to dimension reduction techniques – PCA, which is a linear dimension reduction technique or t-SNE, which is non-linear. If we use linear dimension technique, we can face the **crowding** problem. In linear dimension reduction, all the high dimensional datapoints are projected linearly, which can fail to preserve the original distance between the datapoints and lead to crowding in lower dimensions or dissimilar datapoints being represented on top of each other.

In t-SNE, data points are projected non-linearly in such a way that distances between data points in lower dimensions reflect the distances in higher dimensions, hence solving the crowding problem.

**For Drink dataset, if all the variables (except density) are projected on to lower dimension (2D), there no natural clusters found. Hence diving the dataset into subsets and trying to find clusters/sub-clusters.**

**Subset Divisions:**

**Subset 1 – fixed acidity, volatile acidity, citric acid**

**Subset 2 – fixed acidity, volatile acidity, citric acid, pH**

**Subset 3 – fixed acidity, volatile acidity, citric acid, pH, residual sugar, alcohol**

**Subset 4 – free sulfur dioxide, total sulfur dioxide, sulphates, chlorides**

**Subset 5 – fixed acidity, volatile acidity, citric acid, pH, residual sugar, alcohol, free sulfur dioxide, total sulfur dioxide, sulphates, chlorides**

**Subset 6 (final subset1) – fixed acidity, volatile acidity, citric acid, pH, alcohol, free sulfur dioxide, total sulfur dioxide**

**Subset 7 (final subset2) – residual sugar, chlorides, sulphates**

## K-Means Implementation

K-Means is a unsupervised machine learning algorithm which groups similar data points together to form groups (clusters). K refers to the number of centre points for these groupings or essentially the number of clusters. For a labelled dataset, K usually becomes the number of classes or categories in the target variable but for unlabelled dataset as ‘Drink’, first objective is to find the appropriate number of clusters.

To find the number of clusters, **Elbow Plot** is used.

Elbow Plot is constructed by plotting inertia for a set of k-values. Range for K values selected is 1-10. Good clustering could be achieved by finding the most significant slope change in inertia for the smallest k value i.e. the point after which the inertia starts decreasing linearly.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

These are two Elbow Plots of the final subsets in Drink dataset. For both plots, the appropriate number of clusters can be 2 or 3. Hence, if the dataset or subset is unlabelled, the number of clusters is almost always a guess (an educated guess if using Elbow plot).

**K Means and t-SNE:**

**For t-SNE visualisation we are using K-Means clusters for colouring. The appropriate number of clusters is selected first, then the K Means fitting is done on the subset. Since the labels are not known for the subsets, to represent clusters in visualisation, K Means is used for colouring so that the data points in the cluster created by K Means can be represented by the same colour.**

## t-SNE Tuning

Perplexity: In simple terms, perplexity hyperparameter in t-SNE defines the effective number of neighbours for each datapoint. It can be viewed as a knob that sets the number of effective nearest neighbours. Optimal value of perplexity depends on the amount of data available. Suggested range is 5-50.

Number of Iterations: Number of times the algorithm needs to run to get the optimal results (patterns). Min value is 250.

For each subset hyperparameter values are following:

|  |  |  |
| --- | --- | --- |
| Subset | Perplexity | Number of Iterations |
| Subset1 | 25 | 3000 |
| Subset2 | 30 | 4000 |
| Subset3 | 50 | 4000 |
| Subset4 | 35 | 5000 |
| Subset5 | 40 | 6000 |
| Subset6 (final subset 1) | 50 | 6000 |
| Subset7 (final subset 2) | 30 | 5000 |

## Cluster Interpretations

The dataset, as a whole, didn’t have any explainable patterns. But dividing it into subsets and keep building the subsets to include more variables resulted in two final subsets with labels.

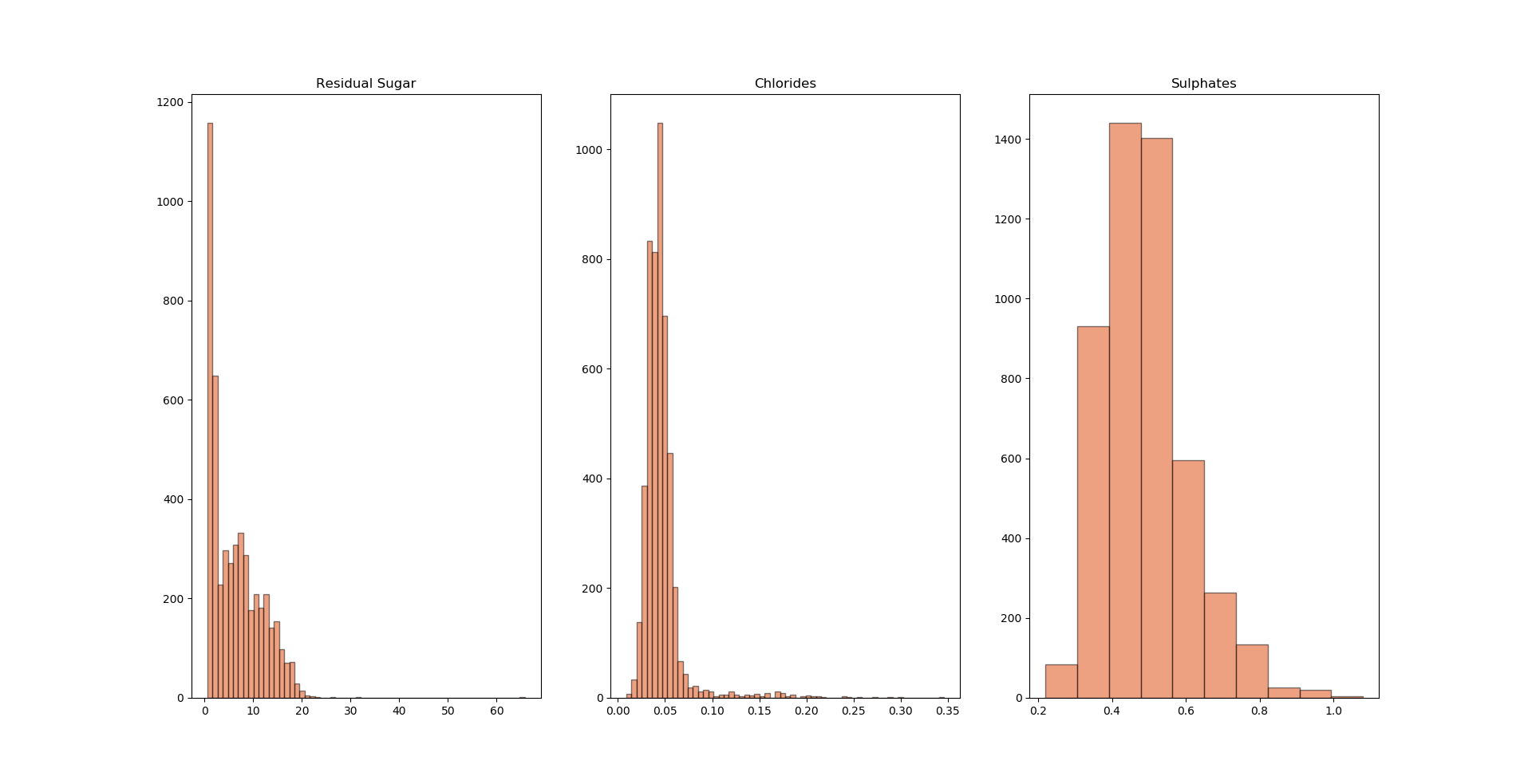
Final Subset 1: fixed acidity, volatile acidity, citric acid, pH, alcohol, free sulfur dioxide, total sulfur dioxide

Final Subset 2: residual sugar, chlorides, sulphates

**After seeing explainable clusters, we needed to convert continuous variables into discrete/categorical label. For that, using bar plots to see the distribution for each variable –**

**A close up of a map

Description automatically generated**

****

**A picture containing boat

Description automatically generated**

**A map of a building

Description automatically generated**

## Final Subset 1 (subset6 in code) Interpretation:

Final Subset 1: fixed acidity, volatile acidity, citric acid, pH, alcohol, free sulfur dioxide, total sulfur dioxide – t-SNE HTML File **– Viz1\_Tanuj\_Gupta.html**

To create this subset, I started with three variables – fixed acidity, volatile acidity and citric acid. I got two clear clusters formed for these variables.

Since ‘fixed acidity’ have a normal distribution, and the patterns for this small subset was clearly divided by the mean. Defined a converter function to check if fixed acidity <= mean thenn defining it as low fixed acidity and if fixed acidity> mean then high fixed acidity.

**I had found a small subset with fixed acidity as label. To further analyse the dataset, I tried to add more variables in the dataset to see if the pattern holds.**

**Finally, I had found a subset which not only holds fixed acidity as label, it has sub-clusters with two more labels.**

**LABELS FOR FINAL SUBSET 1:**

**‘fixed acidity’, ‘alcohol’ and ‘free sulfur dioxide’**

Since all three variables follow normal distribution, all the labels are categorised binarily.

**Interpretation:**

**In the t-SNE plot, we see 8 different sub-clusters for this subset. All the sub-clusters can be labelled with these three labels.**

**Patterns in each sub-cluster:**

Fixed Acidity= High, Alcohol= Low, Free Sulfur Dioxide = Low

Fixed Acidity= High, Alcohol= High, Free Sulfur Dioxide = Low

Fixed Acidity= Low, Alcohol= Low, Free Sulfur Dioxide = Low

Fixed Acidity= Low, Alcohol= High, Free Sulfur Dioxide = Low

Fixed Acidity= Low, Alcohol= Low, Free Sulfur Dioxide = High

Fixed Acidity= Low, Alcohol= High, Free Sulfur Dioxide = High

Fixed Acidity= High, Alcohol= Low, Free Sulfur Dioxide = High

Fixed Acidity= High, Alcohol= High, Free Sulfur Dioxide = High

**To summarise, this subset can be used to predict fixed acidity or alcohol or free sulfur dioxide levels by keeping all other variables as input variables.**



This tableau file shows the variations of citric acid, volatile acidity , pH and total sulfur dioxide with respect to these label filters.

## Final Subset 2 (subset7 in code) Interpretation:

Final Subset 2 – residual sugar, chlorides and sulphates – t-SNE HTML File **– Viz2\_Tanuj\_Gupta.html**

This subset forms three sub-clusters. Since Residual sugar variable does not have a normal distribution (refer chart above) and it more skewed towards right, dividing the variable in three categories.

**Low Sugar: ‘residual sugar’ <=1.7 (1st quartile or 25%)**

**Medium Sugar: ‘residual sugar’ > 1.7 and ‘residual sugar’ <=5.2 (average)**

**High Sugar: ‘residual sugar’ >5.2**

This subset clearly has three sub-clusters. Labelling is done using residual sugar variable. This subset can be used to predict residual sugar levels by using chlorides and sulphates as input variables. Both Chlorides and Sulphates follow the same patterns (increasing/decreasing) for all three levels of sugar.



This tableau file shows variations of chlorides and sulphates with respect to all three levels of sugar.