# Problem Statement

A wine has a lot properties which determines the quality of wine. However, for a layman, it is difficult to judge the quality of wine based on parameters such as acidity, sugar etc. Canada’s wine industry creates about $6.8 billion economic impact and is a significant driver to Canadian economy. More than 31000 jobs in Canada are related to wine industry. Canadians enjoy more than 1 billion glasses of wine, therefore, as a consumer it is important to know the quality of wine. Since, it is hard to know what makes a good wine, therefore the quality can be judged by applying Machine Learning algorithm to various factors.

# Scope of the project

The scope is limited to predictions only of wine quality and no other beverage. Once the best model is selected and it performs well on the testing data, the final model and the code will be the deliverable.

# Dataset

For this problem, I will be using Wine Quality Dataset from Kaggle, that will help use to predict whether the quality of wine is good or not. The variables for predicting the wine quality are-

### Independent variables

* Fixed acidity
* Volatile acidity
* Citric acid
* Residual sugar
* Chlorides
* Free Sulphur dioxide
* Total Sulphur Dioxide
* Density
* Ph
* Sulphates

### Dependent variables

* Quality (1-10)

The dataset has about 1600 rows and 12 columns.

Since it will be hard for regression as well as classification algorithm to predict 10 different outputs just with a small amount of data, therefore we will manipulate the dataset and change the output variables by changing the output into categorical variables as follows-

* The score between 1 and 4 will be modified to **poor.**
* The score between 5 and 7 will be modified to **normal**.
* The score between 8 and 10 will be modified to **good**.

Therefore, with the three unique outputs rather than 10, it will be easy for and algorithm to predict the quality of the wine.

# Dataset Source

The source of the dataset is-

<https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>

# Assumptions

It is assumed that the dataset is from authentic sources. Therefore, no changes will be made to the original values of the dataset. There seems to be the sufficient variables in the dataset, therefore we assume that there is no default assumptions to be made.

# Tasks done

1. Get the dataset
2. EDA
3. Visualizations
4. Scaling
5. Feature Engineering

## EDA

* Display number of rows and columns
* Display key statistics
* Display unique dependent values
* Check dataset through displaying first five columns

## Visualizations

* Display relationship between Fixed acidity vs Volatile acidity
* Heatmap to get the correlation ships
* Display count of each variable
* Display box plots
* Display pair plot