Materials and Methods:

The data set used for this analysis is available at UCI repository website for research purposes under the URL <https://archive.ics.uci.edu/ml/datasets/Wine+Quality> (reference). This data is based on Portuguese “Vinho Verde” red wine. Vinho verde is a wine that is produced in the northwestern region of Portugal called Minho. This wine accounts for 15% of the wine production in Portugal. In this data analysis only red wine has been selected and the white wine variant of vinho verde has been excluded (Corteza, Cerdeira, Almeida, Matos, & Reis, 2009).

A large dataset of red wine comprised of 1599 samples have been utilised, in which 11 independent variables and 1 dependent variable, which is quality have been considered in the data analysis. The independent variables include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH values, sulphates and alcohol (Corteza, Cerdeira, Almeida, Matos, & Reis, 2009). Among the aforementioned 11 independent variables, alcohol and volatile acidity have a profound effect on the aroma and flavour of wine (Reference). Alcohol pertains to the product of fermentation created by the yeast enzymes that convert the grape sugars into ethanol and carbon dioxide. Volatile acidity is a by product of the oxidation of alcohol to acetic acid. It enhances the wine aroma in small amount; whereas, it causes vinegary taste and odour in excessive quantity (Grainger & Tattersall, 2016). The acidity of wine could be determined in terms of pH value. The pH value of 0 is termed as acidic, while pH value of 14 indicates that the wine is alkaline. On average, wine has a pH value of 3 to 4 (reference).

The aim of this analysis is to predict whether the percentage content of alcohol present in a wine affects its quality or not. Due to confidentiality issues, the brand name of wines, its selling price, type of grapes used and wineries locations are not provided in this dataset. The table below provides the mean and standard deviations of all 11 independent physiochemical properties of red wine.

**Data mining approach & techniques:**

Statistical machine learning is a set of powerful algorithms used to predict an outcome variable against multiple predictors. The algorithms improve their performance automatically through learning from the data. The data mining has been divided into two parts: supervised learning and unsupervised learning. Supervised learning captures and processes the required input and conveys output precision and accuracy for the majority of the decision-making problems (Haney, 2015). In supervised learning, one quantitative analysis and one qualitative analysis have been performed. Descriptive statistics has been utilised for quantitative analysis to obtain the mean, median and minimum vallues for each independent variable. In addition, a box plot has been used to indicate the presentation of different parameters of the data (Bruce & Andrew, 2017).

To assess the performance of alcohol in wine using the aforementioned dataset, several tests have been conducted such as multiple linear regression and cross validation approach. Multiple linear regression has been performed to test the correlation between various independent variables and for the dependent variable, which is quality. Multiple linear regression pertains to an extension of simple linear regression in which more than one predictor variables are used to predict an output variable (Bruce & Andrew, 2017). A cross-validation approach has been utilised using alcohol as the only predictor to measure the performance of the resulted model on new test data set. This technique involves dividing the dataset into two subsets known as training dataset and testing dataset. Training dataset is used to train the model while testing dataset is used to test the model by determining the prediction error. There are various cross validation methods to evaluate the performance of a model, such as: validation dataset approach, leave one out cross validation, k-fold cross validation and repeated k-fold cross validation. For this model, the basic validation dataset approach has been used (Bruce & Andrew, 2017).

For this qualitative analysis, decision trees model has been specifically used. It builds a set of decision rules to describe the connection between independent variables and the outcome variable (P. Bruce and Bruce 2017).

For unsupervised learning, Principal Component Analysis has been performed. This method provides a summary and visualization of the information contained in a dataset having multiple quantitative variables. Each component in the dataset is considered a different dimension; however, if there are more than 3 predictor variables in the dataset, it becomes difficult to visualize a multi-dimensional hyperspace. In this situation, PCA extracts the important information from a multivariate data and expresses it as a new set of fewer variables which are known as principal components (PCs). Newly produced variables correspond to a linear combination of the original variables. The number of PCs is less than or equal to the quantity of original variables. In essence, Principal Component Analysis reduces the dimensions of a data to two or three PCs. Therefore, the data could be visualized graphically without any major losses of data information (Reference).

**References:**

Bruce, Peter, and Andrew Bruce. 2017. *Practical Statistics for Data Scientists*. O’Reilly Media.

**Additional References:**

Corteza, P., Cerdeira, A., Almeida, F., Matos, T., & Reisa, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems, 47*, (4). <https://doi.org/10.1016/j.dss.2009.05.016>

Grainger, K. & Tattersall, H. *Wine Production and Quality.* John Wiley & Sons, Incorporated, 2016. Retrieved from <http://ebookcentral.proquest.com/lib/acu/detail.action?docID=4306430>

Haney, S. A. (2015). *An Introduction to High Content Screening : Imaging Technology, Assay Development, and Data Analysis in Biology and Drug Discovery*. Retrieved from <http://ebookcentral.proquest.com/lib/acu/detail.action?docID=1895676>.