|  |  |
| --- | --- |
|  | |
| Python Machine Learning Project  Data and Web Mining CA – Report | |
|  | |
| Ciaran Finnegan / Dermot Madsen  Student No : 10524150 / 10522567  10/02/2020 |  |
|  |  |

List of contents

1 Project Overview 3

1.1 High Level Description 3

1.2 The CRISP-DM Methodology / Reference Model 3

1.3 Development Environments 6

2 Business Understanding 7

2.1 Determine Business Objectives 7

2.2 Assess Situation 8

2.3 Determine Data Mining Goals 9

2.4 Produce Project Plan 9

3 Data Understanding 11

3.1 Collect Initial Data 11

3.2 Describe Data 12

3.3 Explore Data 13

3.4 Verify Data Quality 22

4 Data Preparation 24

4.1 Select Data 24

4.2 Clean Data 24

4.3 Construct Data 24

4.4 Integrate Data 24

4.5 Format Data 24

5 Modelling 25

5.1 Select Modelling Technique 25

5.2 Generate Test Design 25

5.3 Build Model 25

5.4 Assess Model 25

6 Evaluation 26

6.1 Evaluate Results 26

6.2 Review Process 26

6.3 Determine Next Steps 26

7 Deployment 27

7.1 Plan Deployment 27

7.2 Plan Monitoring and Maintenance 27

7.3 Produce Final Report 27

7.4 Review Project 27

8 Conclusion 28

8.1 Conclusion.. 28

9 Appendices and References 29

9.1 Appendix A Understanding Wine and Types 29

9.2 References 31

# Project Overview

## High Level Description

This document covers our planned approach and execution of a data mining analysis on a dataset relating to the assessment and prediction of wine quality.

Following the CRISP-DM model, we laid out an objective for this Continuous Assessment exercise and followed a series of steps, often iteratively to arrive at a predictive model for wine quality based on known feature characteristics.

The following sections explain the business objectives, the assessment of data, and the selection, implementation and deployment of a model to provide a predictive guide to new wine quality.

## The CRISP-DM Methodology / Reference Model

In the mid to late 1990s, business markets were showing a sharp upturn in interest into the possibilities offered by data mining practices. The need for a standard process model, widely and freely available, became quickly apparent.

By 1999/2000, a process model named CRISP-DM (Cross-Industry Standard Process for Data Mining) had been produced by leading thinkers in the industry. It was based on practical, real-world experiences and sought input across a range of business domains.

As explained in the following sections of this document, we have taken the key principles of CRISP-DM to implement our CA project.

### Methodology

The CRISP-DM methodology is described as a hierarchical process mode with four levels that transition from the generic to the specific;

1. **Phases** – process blocks consisting of several generic tasks.
2. **Generic Tasks** – so called because they are intended to be robust and stable tasks that can apply in any data mining situation.
3. **Specialised Tasks** – a description as to how the generic tasks should be applied in specific situations. Very often these tasks can be performed in multiple orders and repeated a number of times.
4. **Process Instances** – this is a record of the actions, decisions, and results of an actual data mining engagement.

### Reference Model

The life cycle of a data mining project consists of six phases, as shown in this image below.



Figure 1

The sequence of the phases is not rigid. In our Wine Quality project moving back and forth between different phases was frequently required, as expected.

The outcome of each phase determines which phase, or particular task of a phase, has to be performed next. The arrows indicate the most important and frequent dependencies between phases.

As an example, in our Wine Quality CA we needed to…<provide some actual examples – when we have them…!>

The diagram above displays the following phases.

* Business Understanding
* Data Understanding
* Data Preparation
* Modelling
* Evaluation
* Deployment

The next sections of this document elaborate on these phases a little further and the remainder of the report describes the actual implementation of the CRISP-DM against our Wine Quality project.

### Business Understanding

Understand the project objectives and the requirements from a business perspective.

Convert this knowledge into an actual data mining problem definition, along with a project plan that will provide a framework to deliver the business objectives.

### Data Understanding

Start with initial data collection.

Proceed into activities that provide familiarity with the data, including data quality issues, data insight, and possible sub-sets within the data.

### Data Preparation

Activities to construct the final dataset that will be fed into the modelling too.

Data preparation tasks can be performed in multiple orders and over many interactions.

### Modelling

Select and apply various modelling techniques.

Calibrate parameters to provide optimal values within the model.

Revert to data preparation phase, if necessary.

### Evaluation

A high quality data analysis model has been built.

Assess that the model achieves the business objective.

### Deployment

The knowledge gained by the creation of the model will need to be organised and presented in a way that it can be used by the customer.

It is important for the customer to know up front what actions need to be carried out in order to actually make use of the created models.

## Development Environments

The majority of development took place within the RapidMiner toolkit and screenshots are provided to show the step-by-step working.



Figure 2

Supplementaty data investigation was conducted using a Python project developed in Visual Studio 2019. The structure of the project was modularised around a framework similar to the CRISP-DM model.

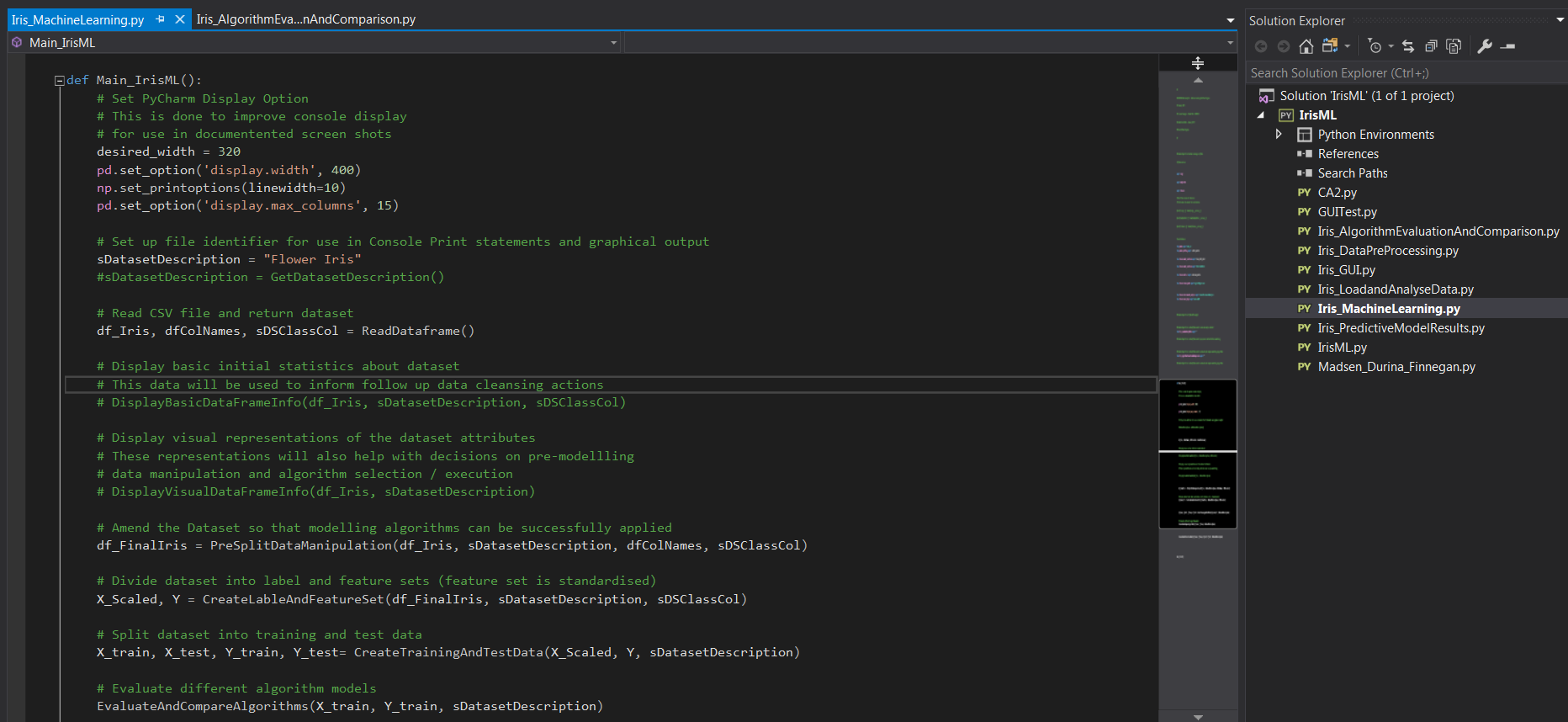


Figure 3

The Python project was used to provide quick to develop validations for the RapidMiner process and the ‘Performance’ outputs provided by that tool.

# Business Understanding

## Determine Business Objectives

The first objective of the data analyst is to thoroughly understand, from a business perspective, what the customer really wants to accomplish.

In our QA, the ‘customer’ and ‘business’ are obviously a theoretical concept. However, given that our chosen dataset relates to Wine Quality, we are assuming the role in the project of a chain of Off-Licence shops, who have a particularly speciality in selling Portuguese “Vinho Verde” red wine. Part of the business USP (unique selling point) is that staff is encouraged to be knowledgeable about the quality of this wine that they may recommend to customers. Although our imaginary Off-Licence chain promotes an awareness of wine amongst staff, very few employees would aspire to the level of sommelier in this brand of Portuguese wine and therefore it is necessary to provide guidance to staff when new stocks of wine arrive in-store.

Vineyards and Wine wholesalers will presumably provide recommendations on what wines are ‘good’ but a secondary objective of our business is to have a less subjective measure of quality for new wines. Our predictive model will therefore provide a more scientific basis for a quality rating, which can be applied across the entire outlet of shops, rather than relying on a human analysis, which could be open to interpretation.

A further secondary objective is that this model may provide guidance for similar in-store marketing of other ‘niche’ wine brands, should our business wish to replicate this approach to wine promotion.

How do we define success? A model is built based to predict the quality of new stocks of red wine based on the constituent chemical properties of the liquid. Our initial dataset will contain information to train and test our model (after various algorithm selections), and we will then conduct an additional test with new ‘unseen’ data to show that the model works well to provide an employee guide to wine quality.

## Assess Situation

This task involves more detailed fact-finding about all of the resources, constraints, assumptions, and other factors to be considered in determining the data analysis goals and project plan.

**Dataset Inventory**

We chose a dataset provided in the Kaggle website ([*https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009*](https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)) that relates to the red wine variant of the Portuguese “Vinho Verde” red wine. (The primary source of the dataset is on the UCI Machine Learning Repository - [*https://archive.ics.uci.edu/ml/datasets/wine+quality*](https://archive.ics.uci.edu/ml/datasets/wine+quality)).

The dataset is publically available and provides a series of input variables, which are based on physiochemical tests, and an output variable that is a 0 – 10 score of quality.

**Assumptions**

The ‘quality’ measure, which is obviously the key characteristic we want to assess, has been assumed to come from feedback from wine industry specialists.

**Constraints**

This an assessment of quality based on the chemical constituents of the wine. There is no data relating to year or grape type, as might be expected with an assessment of wine, so we are assuming that a chemical analysis will provide the all the data points we need for a quantifiable assessment of quality.

There is also no indication of brand or price. This dataset is deliberately excluding these factors (or is unable to include them). Therefore that type of marketing data points will not influence the prediction of ‘quality’ as produced by our model.

## Determine Data Mining Goals

A data mining goal states a project objective in technical terms.

**Goal**

Our CA project aim is to build a predictive model that provides a 0 – 10 rating for a wine based a list of 11 chemical attributes in the liquid.

**Success Criteria**

We want our model to operate with a greater than 85% accuracy in its predictions of wine quality for new “Vinho Verde” red wines.

We may extend the modelling process so that a score of ‘7’ or greater is described as ‘Very Good’, ‘4 – 6’ receives a ‘Good’ description, and anything else is ‘Poor’. Thus we refine our classification of the model outputs into simpler terms for the end user employees.

## Produce Project Plan

**Project Plan**

The framework of this document, even just reading from the Table of Contents, provides the general outline of activity.

In brief, our timelines are to complete the following activity by the following milestones (allowing for some iterations and back and forth before project completion);

Any project plan is a dynamic document, and this CA is no exception and we expected, and encountered, the need for many revisions.

* **Saturday January 25th**: Complete dataset selection and establish business objectives.
* **Saturday February 1st** : Complete Data Understanding, Data Preparation, and preliminary model assessment.
* **Saturday February 8th** : Complete Modelling and Evaluation, determine Production approach. Present to class.
* **Sunday February 9th** : Submit CA final report with recommendations.

**Assessment of Tools**

In order to gain an insight into the use of commonly used industry tool, the majority of the data mining approach was conducted in ***RapidMiner*** (as can be seen in the screenshots used throughout this document).

However, early stage data analysis and some preparation used ***Python*** scripting. This was partially because of familiarity with Python from earlier CA work on the course and also to provide some quick additional verification of the RapidMiner outputs.

Excel was used for part of the validation process on predictions made by the model on new data.

# Data Understanding

## Collect Initial Data

This involves the acquisition of data and loading into our chosen data mining tool kits.

**Initial Data Collection Report**

As described in Section 2.2 of this document the dataset for the CA is taken from the Kaggle website, specifically from the URL : <https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>, which in turn references the original UCI Machine Learning source.



Figure n

In Appendix A of this document there is an brief guide to understanding wine types and composition, with particular relevance to the checmical data points in this dataset. (Source : *https://github.com/dipanjanS*).

Downloading the CSV file from Kaggle is a straightforward exercise and the CSV file itself is just 101 kB.

For preliminary data analysis the CVS file on Red Wine quality loads without issue into RapidMiner.



Figure n

## Describe Data

This involves an examination of the ‘gross’ (or ‘surface’) data and a report on the results.

**Data Description Report**

The file is on a CSV format, and contains 1600 row with 12 attribute columns in the following structure:

***Input variables*** (based on physicochemical tests):  
1 - fixed acidity  
2 - volatile acidity  
3 - citric acid  
4 - residual sugar  
5 - chlorides  
6 - free sulfur dioxide  
7 - total sulfur dioxide  
8 - density  
9 - pH  
10 - sulphates  
11 – alcohol

***Output*** variable (based on sensory data):  
12 - quality (score between 0 and 10)

A surface view in NotePad++ shows the following sample structure;



Figure n

In line with the description of the dataset in Kaggle, the dataset contains header information but the remainder of the dataset is purely numeric.

Given that this is a dataset aimed at relative new comers to the work of Machine Learning, it does not seem likely that there will be any invalid or missing data entries. We would expect this to be borne out in the analysis in the following sections of this document.

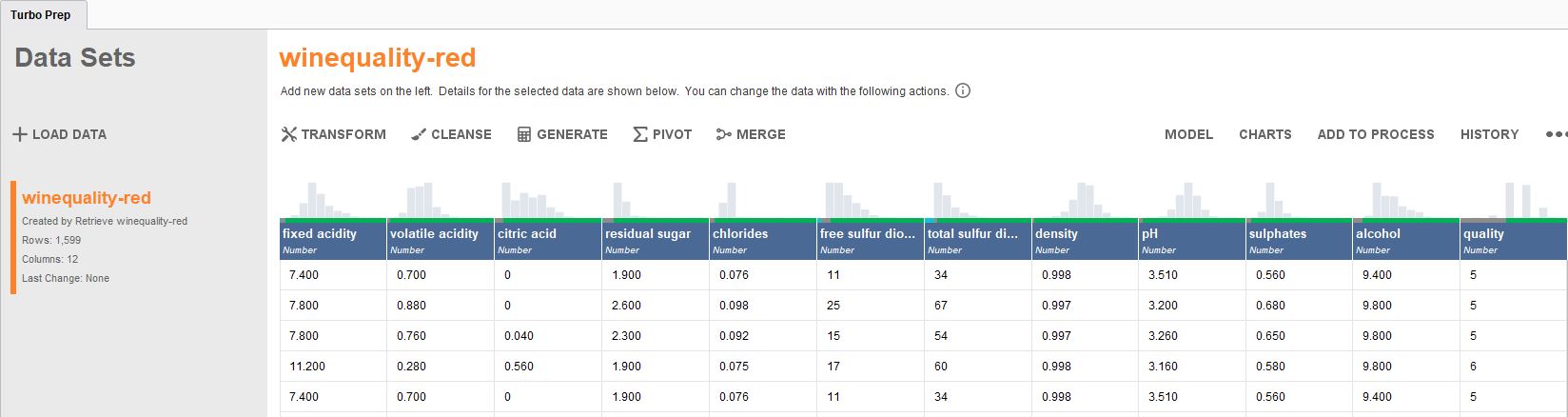
## Explore Data

This task addresses data mining questions using querying, visualization, and reporting techniques.

**Explore ‘Wine Quality’ Data**

*Assessing the Target attribute*.

The figure below shows the head rows in the dataset, in a much more readable format than with Notepad++.



As described eearlier, the are 1600 rows in the dataset (one row with header information), and 12 columns.

The purpose of our data mining task to assess wine quality. Therefore the attribute column for ‘**quality**’ will be marked as our **label**.This is the value which we want to predict, based on the entries in the other other attribute rows.

As will be explained in Section 4 of this CA report, we will carry out feature engineering to add an additional attribute to make the interpretation of the model easier for the end user by moving away from a numeric output to a qualitiative description.

*Relationships between attributes. Correlations of attributes to the label value and between attributes themselves.*

In many datasets not every attribute will have a strong impact on the target label. The values of certian attributes may influence the model very little and can actually be ignored in the model building process.

In addition, certain attributes may be strongly correlation with each other, either negatively or positively. Commonly cited examples in data modelling are if measurements are stored using different standards – for example metric vs imperial. One of those attributes will be redundanct because an increase in one attribute is reflected with a similar scale of increase in the other.

For our wine quality dataset we can execute a correlation analysis on the attributes to determine which attribute have the greatest impact on the resultant quality of the wine.

RapidMiner provides a process to carry out this analysis, either through a stand-alone process using the Correlation Matric Operator, or throught he AutoModelling process.

After conducting such an analysis on our database, we can see the following correlation data between attributes and label;

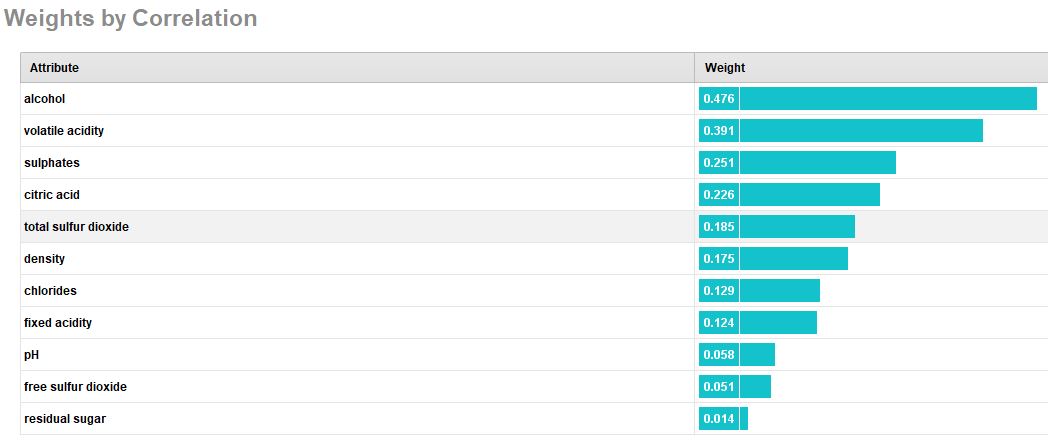


Figure n

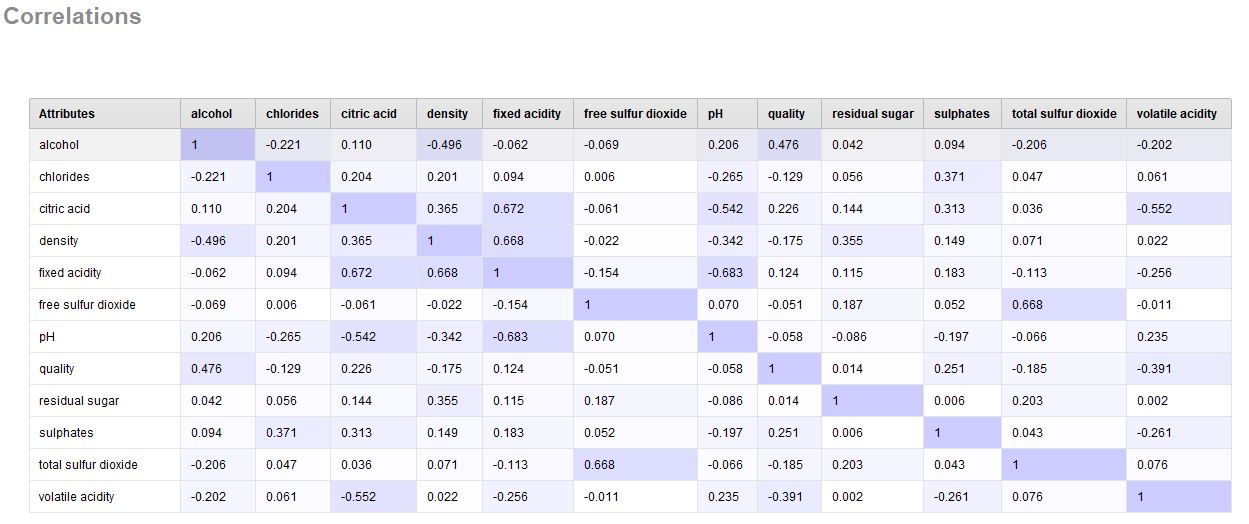
Alcohol content has the greatest impact on the resultant measure of ‘quality’.

Residual sugar appears to have the least impact on the determination of red wine quality in our sample dataset.

The diagram above shows the other relative metrics for each attribute in terms of their influence on ‘quality’.

Within the attributes themselves, correlations can be examined to determine if some attributes can be considered redundant or, at the very least, less important to include in the model calculations.

The diagram below shows the inter attribute correlations;



The darker colours indicate attributes in the feature set that have higher correlation values.

Not surprisingly the **pH** and **fixed acidity** attributes are showing are more markd negative correlation that many of the other features.

Likewise the **free sulfur dioxide** and **total sulfur dioxide** features are showing a reasonably positive correlation.

Although the Wine Quality dataset is not large in terms of rows and attributes, the principle holds that models can be improved in terms of creation and execution if the feature set is pruned to just those attributes that have the greates impact in the model accuracy.

Taking the above considerations into account the list of attributes with which to train our potential models, which is described in detail in Section 5 of this document, can be pruned to improve performance.

*Simple Statistical Analysis of the Wine Quality Dataset*

RapidMiner provide a simple and graphical means to generate some basic statistical data on the attributes in the dataset.

The figures below provide a visual description of the histograms for each attribute in the Wine Quality dataset.

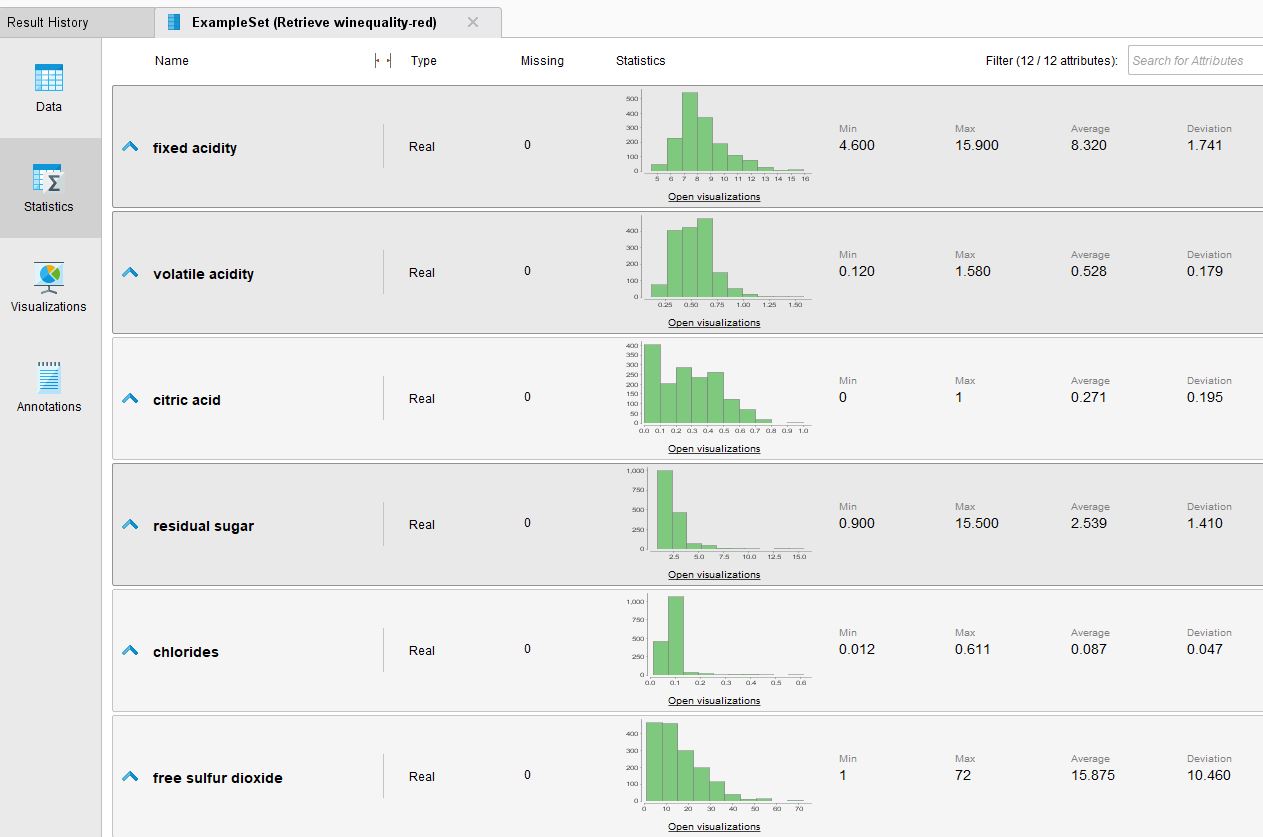


Figure n



Figure n

**Data Exploration Report**

*Initial findings / hypotheses and their impact on the remainder of the Wine Quality project.*

*Standardise and normalise the feature attributes*

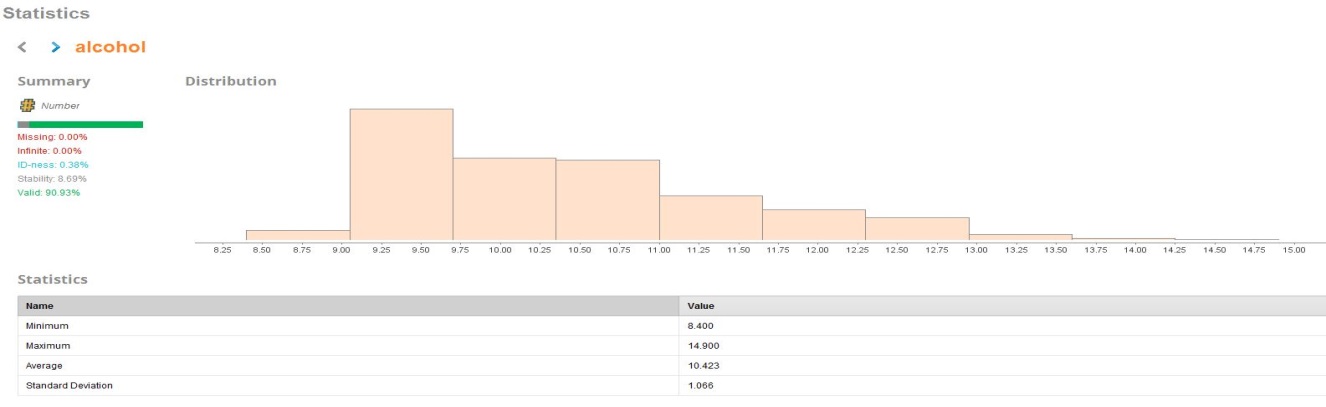
All the values in the dataset are numeric, which will simply some the data preparation tasks as the data mining process needs numeric data to build a mathematical model.

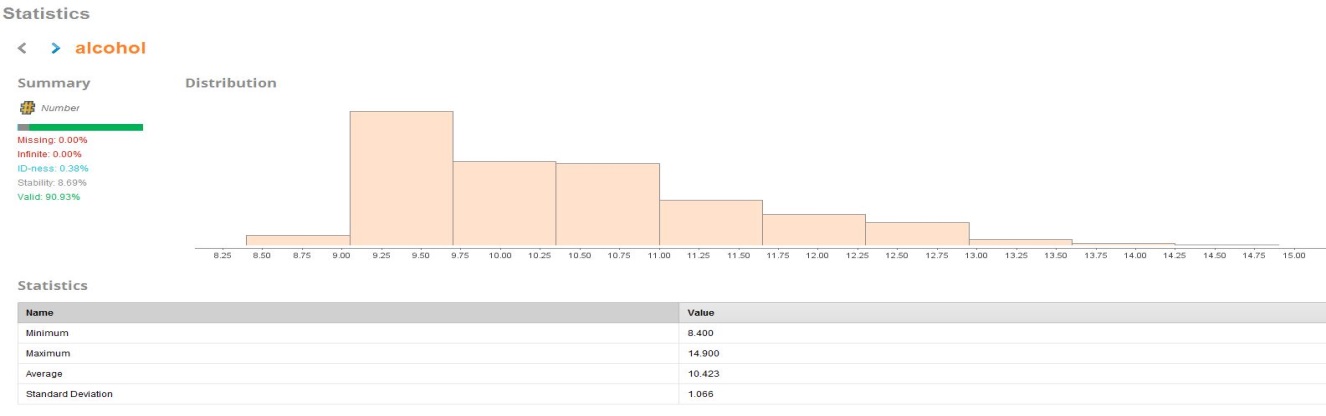
However, although the range of numeric data is not particularly large there are still some features that use a noticeably difference scale, froexample the free sulfur dioxide range of values.

In order to prevent such attributs from skewing the resultant data model, all the elements in the fature set will be normalized. This process is described in more detail in Section 4 of this report.

*Quality of Data – No Missing Rows*

The following three screen shots are from the AutoModel output of Rapid Miner. They are a sample of the ‘General’ output from the AutoModel process.





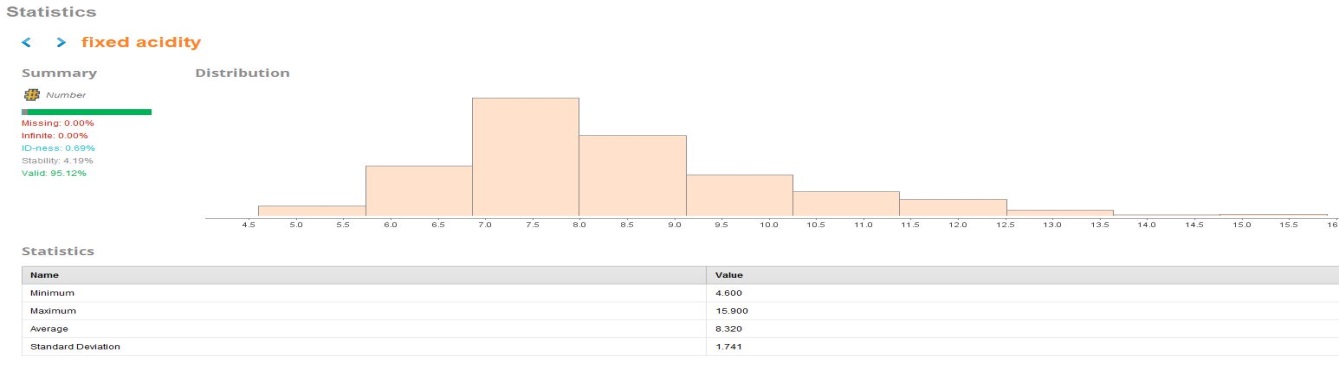


Figure n

The RapidMiner analysis shows that there are no missing data elements in any of the attributes in the dataset.

This will reduce the complexity in the data preparation stage of the project as it will not be necessary to impute missing data, nor will it be necessary to remove incomplete rows from the dataset.

We can see this again in the list of attributes in the ‘Result’s tab for the dataset in RapidMiner.

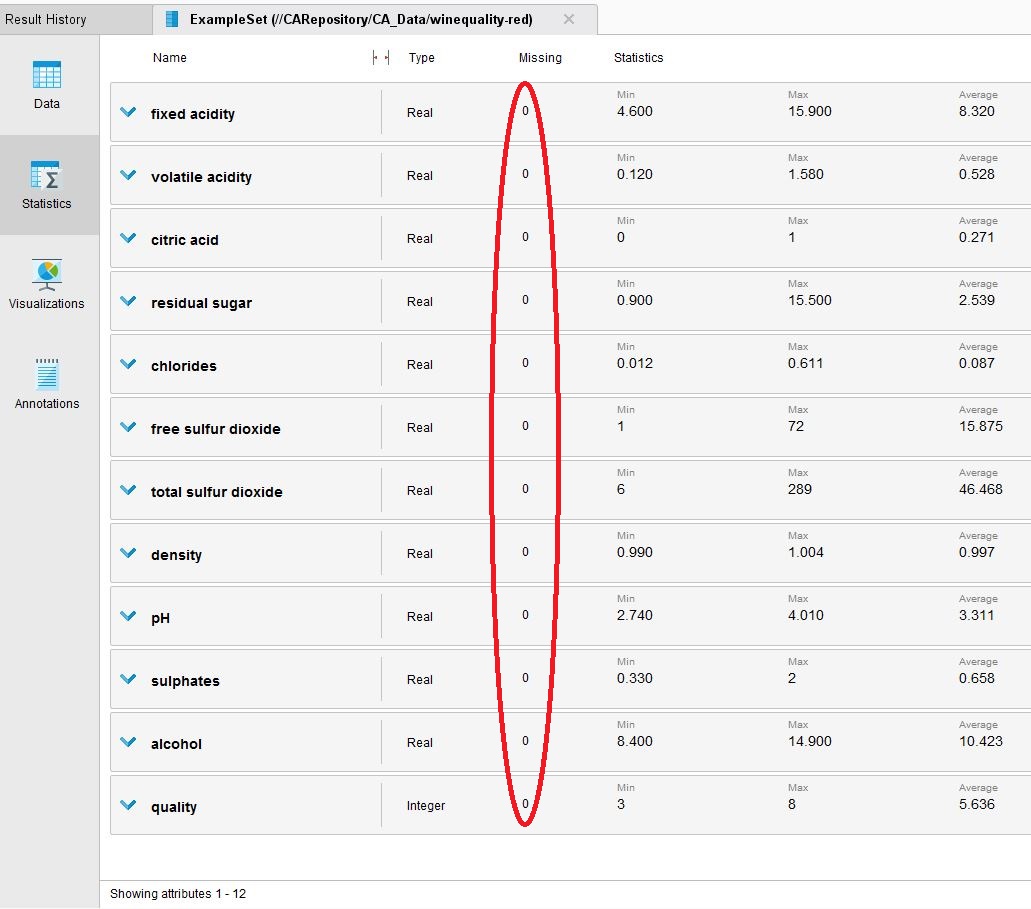


Figure n

*Quality of Data – Duplicate Values*

A quick Python based routine was run independently to confirm that norows in the Wine Quality dataset are duplicates.

<Python screenshot>

*Quality of Data – Zero Values*

However, although our analysis shows no missing rows or obvious errors in the format of data in the columns, we ran an additional check for ‘zero’ values.

A zero entry, particularly when there is a significant proportion of information in a dataset that is numeric format, can also be an indication of missing data.

A quick EXCEL analysis of the Wine Quality csv file will show that the are **132** rows for the ‘citric acid’ attribute that have a zero value.

A quick validation in our supplementary Python program will show a similar analysis.

<Python screenshot>

Is this a legitimate data entry, or are these data rows incomplete?

To investigate further we referenced supporting material on Kaggle, and elsewhere, that explains the fermentation process for wine in more detail. Ctric acid is typically only found in small quantities in wine, and it is used to add what is described as ‘freshness’ to the wine by enhancing flabour.

It is not unusual that the citric acid component in a given type of wine to be completely consumed during the wine fermentation process, and it is not always added back into the process.

Therefore we have concluded that a ‘zero’ entry for citric acid is a valid data point and will not need to be addressed in the data preparation phase of this project.

*Balancing the Wine Quality Dataset*

There is a concern on the spread of wines in the dataset in terms of quality.

RapidMiner allows for a more details bar chart view on the ‘quality’ data in the Wine Quality dataset.

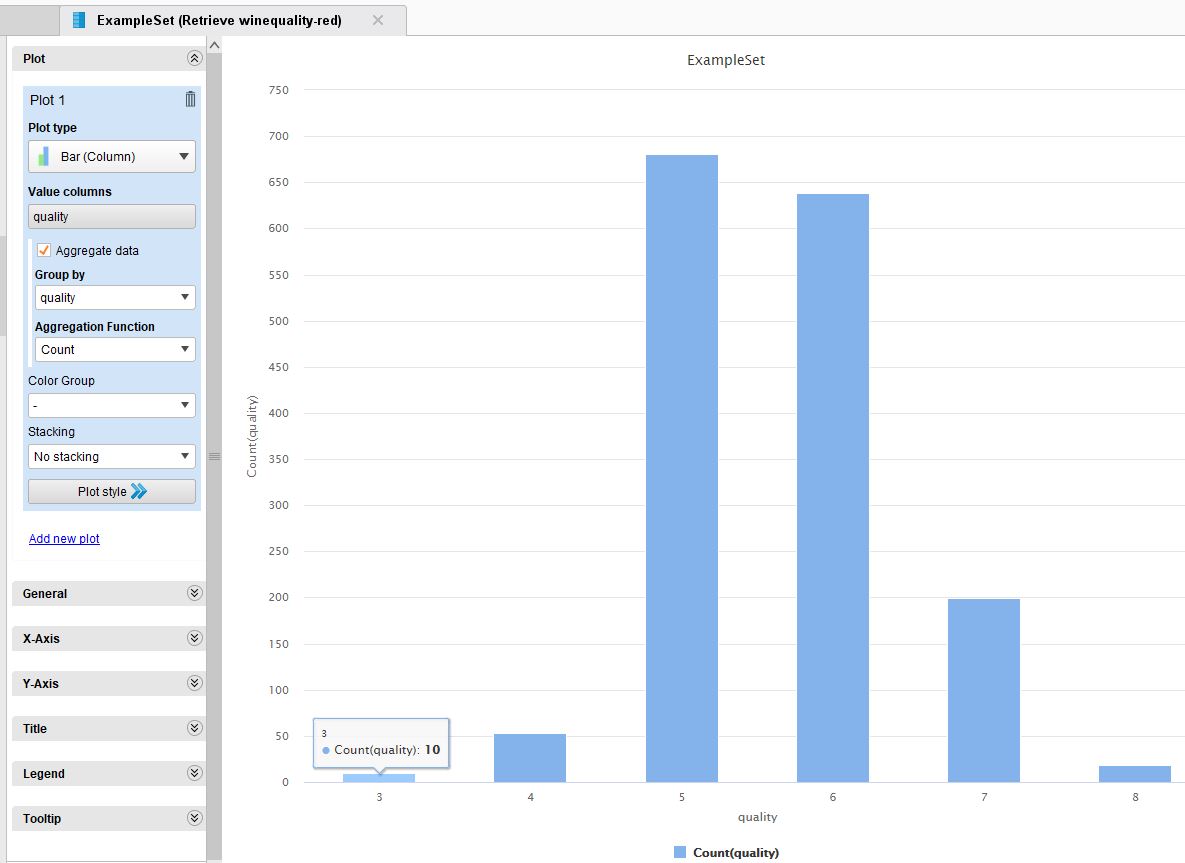


Figure n

There is an immediate concern that the dataset has a high count volume of wines with a ‘quality’ rating of **5** or **6**. However there are relatively few of either the very low or vry high quality red wine types.

The Wine Quality dataset is unbalanced and does not have an even spread of wine qualities. This will impact on the accuracy of any model we attempt ot build.

Section 4 of this report looks at steps we need to take to balance the Wine Quality dataset before we attempt to train and evaluate any models.

The bar chart above provides a good graphical representation of the spread. For a simpler ‘at-a-glance’ view the Python project output at this phase of data exploration shows the following distribution of ‘quality’ classes.

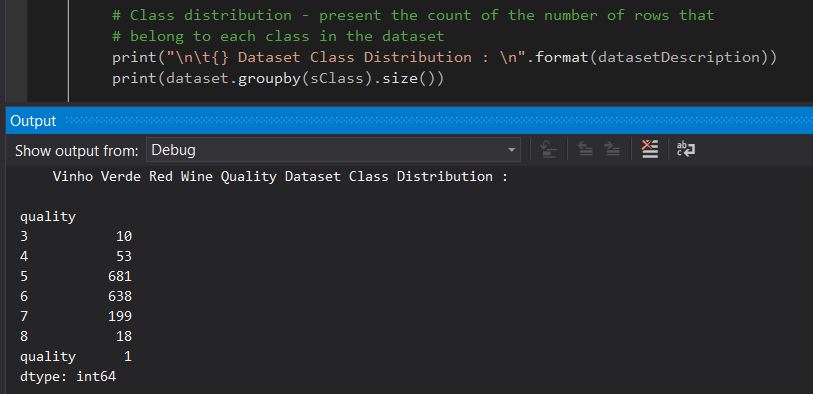


Figure n

## Verify Data Quality

Is the data complete? Does it contain errors and/or missing data? If so, how common are these issues?

**Data Quality Report**

Data quality has been shown to be very good based on the source description and our own data analysis.

There are no missing entries and no obvious errors in the dataset. There are no documented errors in the comments section on the Kaggle page from which the dataset was downloaded and this provides additional assurances.

*Checking for Outliers*

It is impractical to visually assess outliers with the numbers of features in the Wine Quality dataset.

However, RapidMiner provides Operators to detect outliers in datasets.

The following simple DataMiner process was created to use Local Outlier Factors to generate outlier scores for each row in the Wine Quality dataset.

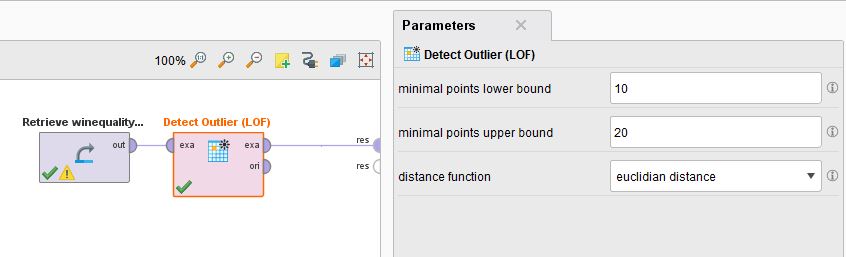


Figure n

The data shows two distinct outliers in the result set.

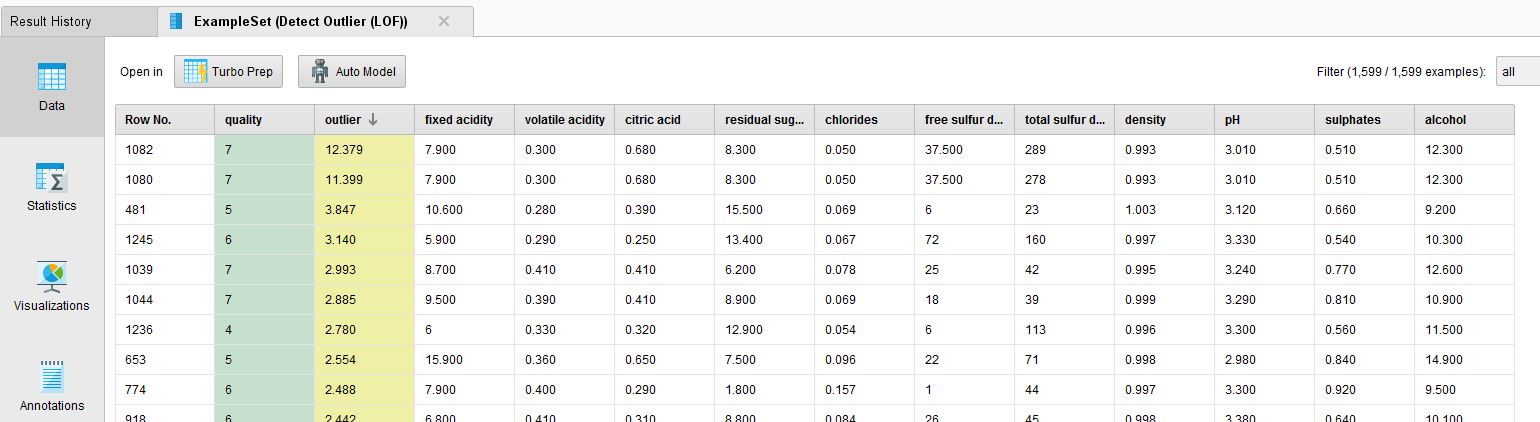


Figure n

A more graphical representation in DataMiner display the outliers in the following Scatter Plot;

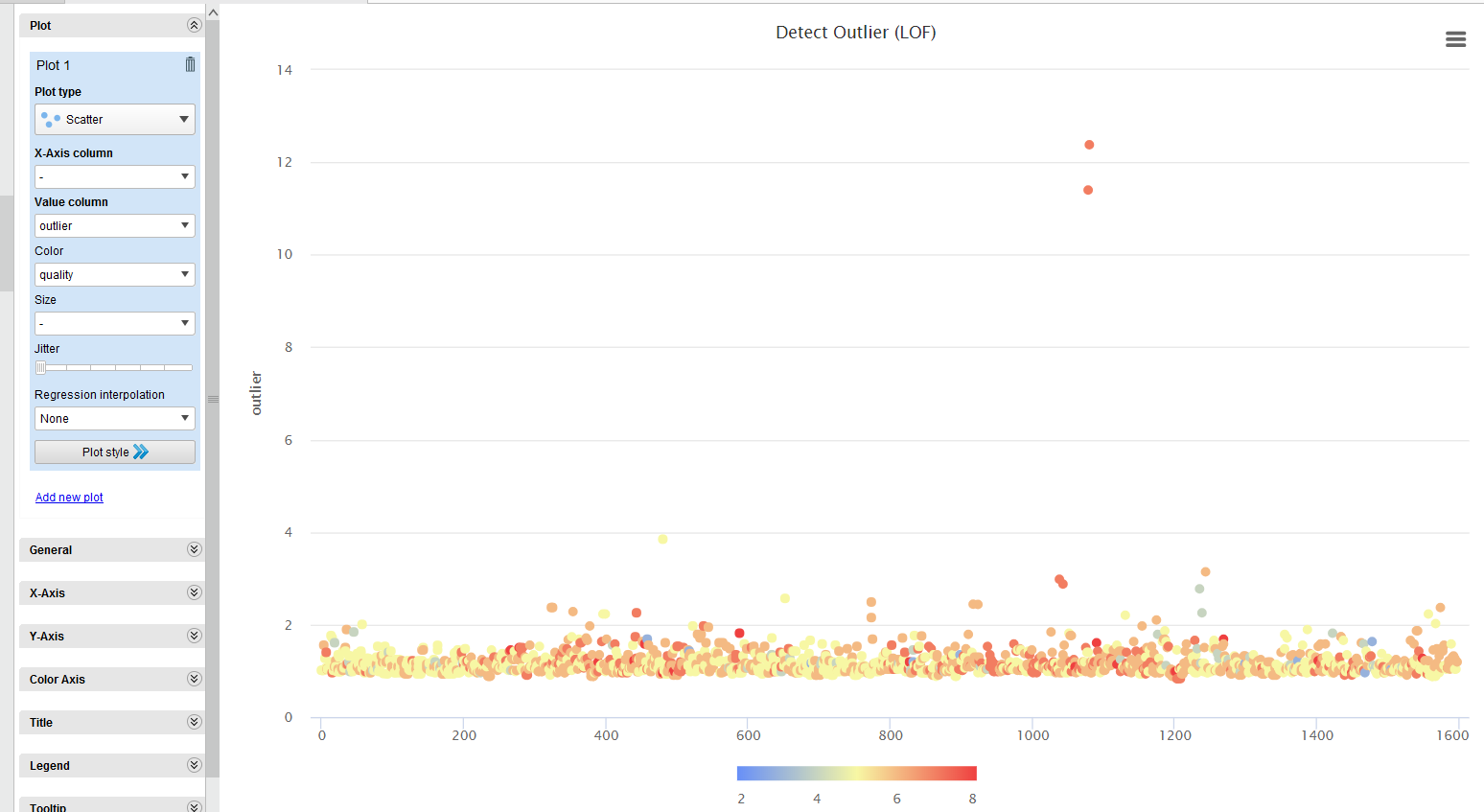


Figure n

A kwy pont to consider is that the ‘outliers’ are marked as ‘high quality’ wines, which can be seem by the colour coding of the entries in the scatter plot above.

As already described, there are very few high quality wines in the dataset hence it is logical that this data would appear as outliers.

Deleting the outliers in this dataset would actually remove key information used in the building of the predictive model. As discussed in Section 4 of this document, it will be necessary to augment the dataset by artificially adding ‘high quality’ data rows to accompany these outliers.

# Data Preparation

## Select Data

The Wine Quality dataset is relatively small so there is no requirement to reduce the number of rows upon which we will build our model in RapidMiner.

There are no quality issues that warrant the removal of any data rows.

AutoModel analysis provided the guideline to a reduced feature set.

<screenshot>

There was a considerable amount of iteration from model selection and evaluation back to feature seelcton to determine the impact of various attribute lists.

## Clean Data

Not an issue with this Kaggle dataset...

## Construct Data

Generate new attribute called ‘wine Quality’ to allow for more meaningful categorization of the data.

Exmplain data range – showscreen shot

Shuffle data to remove potential bias – dataset is described as ordered.

Normalise the attribute values to prevent bias with larger numerical values.

## Integrate Data

No need to merge additonla data sources, although we do carry out a supplementary ‘white win’ analsysis after the main modelling, evalyation and deployment.

## Format Data

Description of upsampling and downsampling – major part of the report...

# Modelling

## Select Modelling Technique

This is a Classification problem so there are acaemdemic guidelines.

Choose to also use the modelling tool provided by RapidMinior at web link..

Screenshot

## 

## Generate Test Design

Training and test sets..

Show screen shot..

SMOTE and Downsampling within Cross Validation..

## Build Model

The..

## Assess Model

The..

# Evaluation

## Evaluate Results

We..

## Review Process

The..

## Determine Next Steps

The..

# Deployment

## Plan Deployment

Our analysis of..

.

## Plan Monitoring and Maintenance

A ...

## Produce Final Report

A ...

## Review Project

A ...

# Conclusion

## Conclusion..

The..

# Appendices and References

## Appendix A Understanding Wine and Types

Wine is an alcoholic beverage made from grapes which is fermented without the addition of sugars, acids, enzymes, water, or other nutrients

Red wine is made from dark red and black grapes. The color usually ranges from various shades of red, brown and violet. This is produced with whole grapes including the skin which adds to the color and flavor of red wines, giving it a rich flavor.

White wine is made from white grapes with no skins or seeds. The color is usually straw-yellow, yellow-green, or yellow-gold. Most white wines have a light and fruity flavor as compared to richer red wines.

**Understanding Wine Attributes and Properties**

* **fixed acidity:** Acids are one of the fundamental properties of wine and contribute greatly to the taste of the wine. Reducing acids significantly might lead to wines tasting flat. Fixed acids include tartaric, malic, citric, and succinic acids which are found in grapes (except succinic).
* **volatile acidity:** These acids are to be distilled out from the wine before completing the production process. It is primarily constituted of acetic acid though other acids like lactic, formic and butyric acids might also be present. Excess of volatile acids are undesirable and lead to unpleasant flavor. In the US, the legal limits of volatile acidity are 1.2 g/L for red table wine and 1.1 g/L for white table wine.
* **citric acid:** This is one of the fixed acids which gives a wine its freshness. Usually most of it is consumed during the fermentation process and sometimes it is added separately to give the wine more freshness.
* **residual sugar:** This typically refers to the natural sugar from grapes which remains after the fermentation process stops, or is stopped.
* **chlorides:** This is usually a major contributor to saltiness in wine.
* **free sulfur dioxide:** This is the part of the sulphur dioxide that when added to a wine is said to be free after the remaining part binds. Winemakers will always try to get the highest proportion of free sulphur to bind. They are also known as sulfites and too much of it is undesirable and gives a pungent odour.
* **total sulfur dioxide:** This is the sum total of the bound and the free sulfur dioxide. This is mainly added to kill harmful bacteria and preserve quality and freshness. There are usually legal limits for sulfur levels in wines and excess of it can even kill good yeast and give out undesirable odour.
* **density:** This can be represented as a comparison of the weight of a specific volume of wine to an equivalent volume of water. It is generally used as a measure of the conversion of sugar to alcohol.
* **pH:** Also known as the potential of hydrogen, this is a numeric scale to specify the acidity or basicity the wine. Fixed acidity contributes the most towards the pH of wines. You might know, solutions with a pH less than 7 are acidic, while solutions with a pH greater than 7 are basic. With a pH of 7, pure water is neutral. Most wines have a pH between 2.9 and 3.9 and are therefore acidic.
* **sulphates:** These are mineral salts containing sulfur. Sulphates are to wine as gluten is to food. They are a regular part of the winemaking around the world and are considered essential. They are connected to the fermentation process and affects the wine aroma and flavor.
* **alcohol:** Wine is an alcoholic beverage. Alcohol is formed as a result of yeast converting sugar during the fermentation process. The percentage of alcohol can vary from wine to wine. Hence it is not a surprise for this attribute to be a part of this dataset. It's usually measured in % vol or alcohol by volume (ABV).
* **quality:** Wine experts graded the wine quality between 0 (very bad) and 10 (very excellent). The eventual quality score is the median of at least three evaluations made by the same wine experts.
* **wine\_type:** Since we originally had two datasets for red and white wine, we introduced this attribute in the final merged dataset which indicates the type of wine for each data point. A wine can either be a 'red' or a 'white' wine. One of the predictive models we will build in this chapter would be such that we can predict the type of wine by looking at other wine attributes.
* **quality\_label:** This is a derived attribute from the quality attribute. We bucket or group wine quality scores into three qualitative buckets namely low, medium and high. Wines with a quality score of 3, 4 & 5 are low quality, scores of 6 & 7 are medium quality and scores of 8 & 9 are high quality wines. We will also build another model in this chapter to predict this wine quality label based on other wine attributes.

## References

The use of this Wine Quality dataset in this CA acknowledges the source publication :

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.