# Wine Production Quality and Satisfaction: Data Cleaning and Visualization

CSCI 5622: Machine Learning

**Author: Samuel Kwon**

Department of Computer Science
University of Colorado Boulder

09/11/22

# Introduction

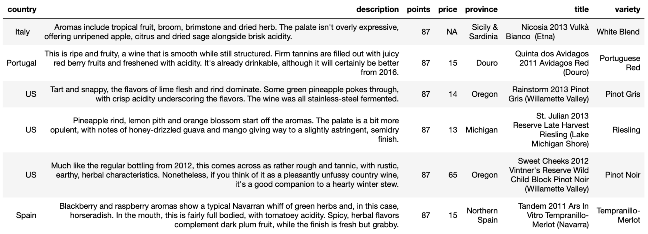
The varieties of red and white wine are increasing every year. The increased production and sales of wine make it harder for the consumer to purchase good quality wine. As a result, the goal of this project is to explore wine datasets and consumer satisfaction. Wine production starts with the harvest of grapes. Different grape varieties grow in various regions. Depending on the variety, the grape can influence the chemical composition, taste of the final product, and overall satisfaction. Consumers' most popular grape varieties are Cabernet Sauvignon, Shiraz, Pinot Noir, Sauvignon Blanc, and Merlot. Not only can the grape variety influence the final product, but the production process plays a role in the quality of the wine. The first step in wine production is the harvest of grapes. Depending on the goal of the final product, one variety or a combination of multiple varieties will be harvested. The combination of multiple varieties is called blended wines, which tend to have their unique chemical composition and profile. The second step in the production process is fermentation by using yeast. Red wine production will include the skin during fermentation and storage. White wines tend to use grapes without the skin, although some practices use the skin contact method. The fermentation process is divided into a two-step process where the first step is yeast fermentation, and the second is malic acid fermentation. During alcohol production and fermentation, wines develop different congeners, influencing the taste and flavor profile. Congeners are organic compounds that influence the flavor profile and quality of the wine. The last step in the production process is aging, where the product is left inside a wooden, glass, clay, or steel container. Just like how whiskey barrels give a wooden, oaky flavor and scent to the final product, aging the wine also changes the alcoholic content and concentration of congener compounds. Wines can also go through additional clarification processes giving the final product. In this machine learning project, wine-related datasets were collected to examine the properties and satisfaction associated with some red wines. Chemical analysis data was cleaned to examine how the chemical composition influences final consumer satisfaction. In addition, wine ratings produced by the Wine Enthusiast were collected to examine relationships between the red wine description and chemical composition with consumer satisfaction. Besides global wine data, wine data from Portugal was also cleaned to examine chemical composition and satisfaction. Lastly, news API data was also included to examine current information about wine available to the public.

# Analyses

Four datasets have been collected and cleaned so far. The first dataset, called the Wine Enthusiast dataset, has qualitative and quantitative variables on consumer reviews. The second dataset is the Chemical Analysis dataset which only has quantitative data on chemical compounds in wine. The third dataset is Portugal Wine data, where the set contains values of chemical compounds and consumer satisfaction. The fourth dataset is News API data showing current news articles related to wine.

## Wine Enthusiast Dataset

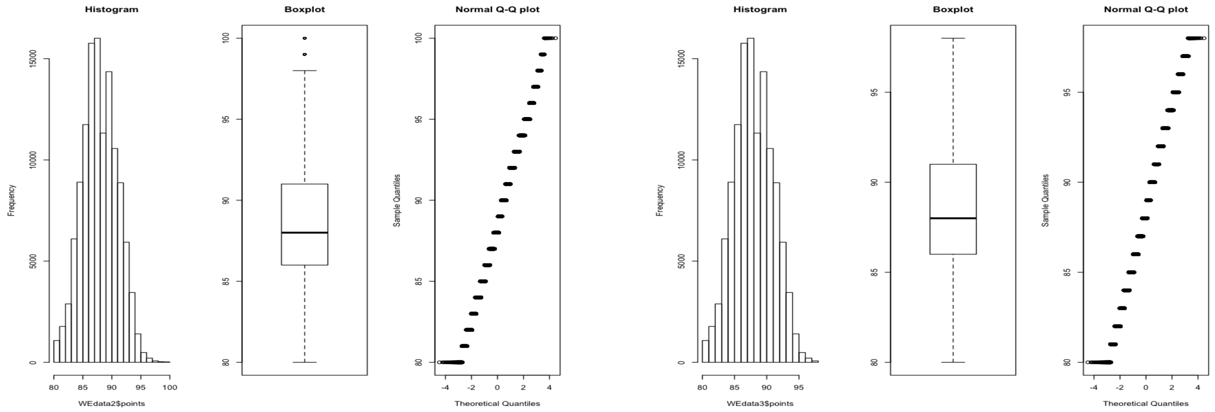
This dataset was obtained through Kaggle, where a user posted web scrapping data from the Wine Enthusiast. The Wine Enthusiast examines independent wines, and the following data values are provided: country, description, points, price, province, title, and variety. The dataset has seven columns and 129971 rows which is a relatively large dataset.



| country | description | points | price | province | title | variety |
|---|---|---|---|---|---|---|
| Italy | Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity. | 87 | NA | Sicily & Sardinia | Nicosia 2013 Vulkà Bianco (Etna) | White Blend |
| Portugal | This is ripe and fruity, a wine that is smooth while still structured. Firm tannins are filled out with juicy red berry fruits and freshened with acidity. It's already drinkable, although it will certainly be better from 2016. | 87 | 15 | Douro | Quinta dos Avidagos 2011 Avidagos Red (Douro) | Portuguese Red |
| US | Tart and snappy, the flavors of lime flesh and rind dominate. Some green pineapple pokes through, with crisp acidity underscoring the flavors. The wine was all stainless-steel fermented. | 87 | 14 | Oregon | Rainstorm 2013 Pinot Gris (Willamette Valley) | Pinot Gris |
| US | Pineapple rind, lemon pith and orange blossom start off the aromas. The palate is a bit more opulent, with notes of honey-drizzled guava and mango giving way to a slightly astringent, semidry finish. | 87 | 13 | Michigan | St. Julian 2013 Reserve Late Harvest Riesling (Lake Michigan Shore) | Riesling |
| US | Much like the regular bottling from 2012, this comes across as rather rough and tannic, with rustic, earthy, herbal characteristics. Nonetheless, if you think of it as a pleasantly unfussy country wine, it's a good companion to a hearty winter stew. | 87 | 65 | Oregon | Sweet Cheeks 2012 Vintner's Reserve Wild Child Block Pinot Noir (Willamette Valley) | Pinot Noir |
| Spain | Blackberry and raspberry aromas show a typical Navarran whiff of green herbs and, in this case, horseradish. In the mouth, this is fairly full bodied, with tomatoey acidity. Spicy, herbal flavors complement dark plum fruit, while the finish is fresh but grabby. | 87 | 15 | Northern Spain | Tandem 2011 Ars In Vitro Tempranillo-Merlot (Navarra) | Tempranillo-Merlot |

Figure 1: Wine Enthusiast Uncleaned Dataset

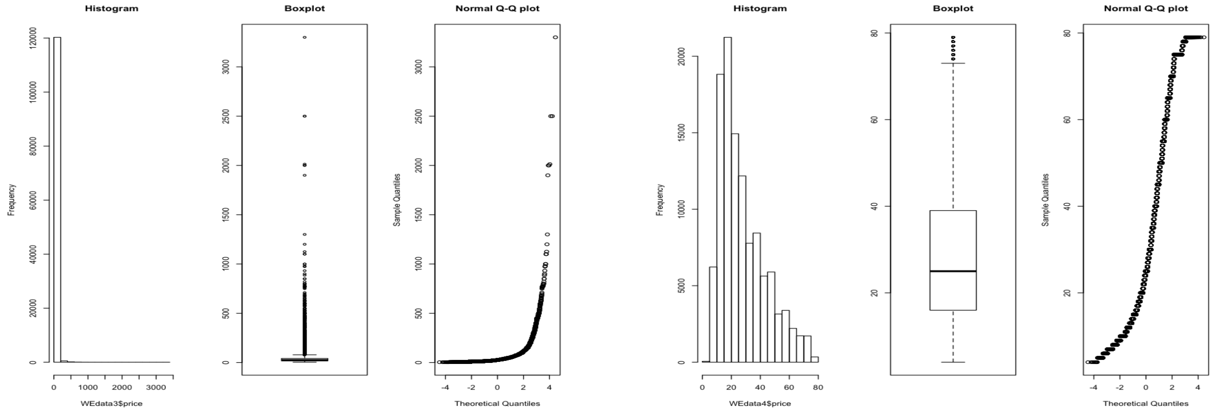
Figure 1 shows that the price and point vectors are incomplete and have some empty values. It is hard to replace wine price data points as they vary drastically and have no standard in determining price. The range of the price dataset is large, meaning that replacing it with mean or median would impact the interpretation and proper representation of the dataset. It is important to note that the dataset has 129953 total rows of data, and only 8996 are incomplete. As a result, the whole row with missing price and points data values was removed from the dataset. This dataset is a combination of qualitative and quantitative values. The price and point data set were examined individually using a histogram, boxplot, and QQ plot. The points histogram, boxplot, and QQ plot are presented in figure 2a. Note that the dataset presents some outliers seen in the boxplot and QQ plot. Because the dataset is relatively large and has some extreme values, the interquartile range method was utilized to remove outliers. Figure 2b is the cleaned points dataset. Notice that the boxplot and QQ plot show that the outliers were removed. The price dataset was also cleaned using the same method. It is important to note that price has extreme outliers since wine prices can vary drastically. Figure 3a and 3b show the uncleaned and cleaned dataset. The cleaned dataset has some outliers shown in the boxplot and QQ plot. The interquartile range method could not remove all the outliers; however, those remaining outliers will be kept as noise for the model. Lastly, the country, description, province, title, and variety data were cleaned by converting the capital words to lowercase letters for text data usage. Figure 4 represents a sample of the cleaned dataset. The final dataset has around 113730 rows, which show a reduction from the earlier dataset.



(a) Uncleaned Points Dataset

(b) Cleaned Points Dataset

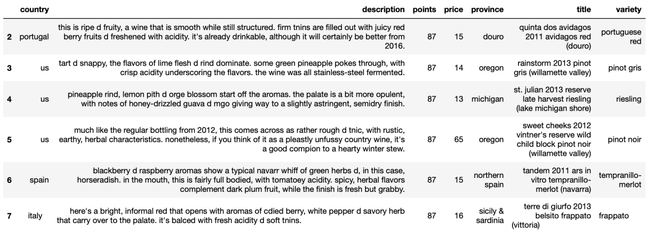
Figure 2: Points Dataset



(a) Uncleaned Price Dataset

(b) Cleaned Price Dataset

Figure 3: Price Dataset

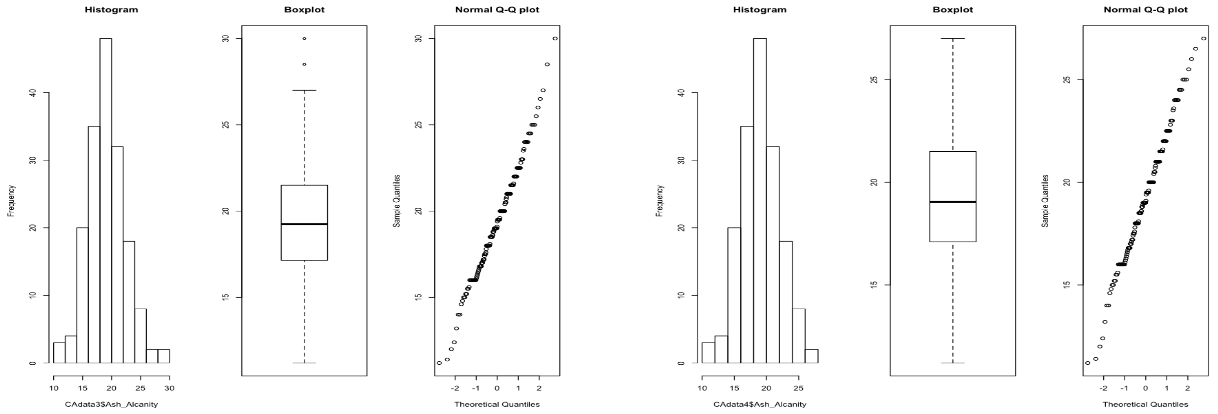


| | country | description | points | price | province | title | variety |
|---|---|---|---|---|---|---|---|
| 2 | portugal | this is ripe d fruity, a wine that is smooth while still structured. firm tnins are filled out with juicy red berry fruits d freshened with acidity. it's already drinkable, although it will certainly be better from 2016. | 87 | 15 | douro | quinta dos avidagos 2011 avidagos red (douro) | portuguese red |
| 3 | us | tart d snappy, the flavors of lime flesh d rind dominate. some green pineapple pokes through, with crisp acidity underscoring the flavors. the wine was all stainless-steel fermented. | 87 | 14 | oregon | rainstorm 2013 pinot gris (willamette valley) | pinot gris |
| 4 | us | pineapple rind, lemon pith d orge blossom start off the aromas. the palate is a bit more opulent, with notes of honey-drizzled guava d mgo giving way to a slightly astringent, semidry finish. | 87 | 13 | michigan | st. julian 2013 reserve late harvest riesling (lake michigan shore) | riesling |
| 5 | us | much like the regular bottling from 2012, this comes across as rather rough d tnic, with rustic, earthy, herbal characteristics. nonetheless, if you think of it as a pleastly unfussy country wine, it's a good compion to a hearty winter stew. | 87 | 65 | oregon | sweet cheeks 2012 vintner's reserve wild child block pinot noir (willamette valley) | pinot noir |
| 6 | spain | blackberry d raspberry aromas show a typical navarr whiff of green herbs d, in this case, horseradish. in the mouth, this is fairly full bodied, with tomatoey acidity. spicy, herbal flavors complement dark plum fruit, while the finish is fresh but grabby. | 87 | 15 | northern spain | tandem 2011 ars in vitro tempranillo-merlot (navarra) | tempranillo-merlot |
| 7 | italy | here's a bright, informal red that opens with aromas of cdied berry, white pepper d savory herb that carry over to the palate. it's balced with fresh acidity d soft tnins. | 87 | 16 | sicily & sardinia | terre di giurfo 2013 belsito frappato (vittoria) | frappato |

Figure 4: Wine Enthusiast Cleaned Dataset

## Chemical Analysis Dataset

The Chemical Analysis dataset is from the UCI Machine Learning repository. This dataset results from a chemical analysis of wines grown in Italy. Around 13 column variables are present, and they are Alcohol, Malic Acid, Ash, Ash Alcanity, Magnesium, Total Phenols, Flavanoids, Nonflavanoid Phenols, Proanthocyanins, Color Intensity, Hue, OD280, and Proline. The dataset has 178 rows, and no values are missing from the set. Since the dataset is quantitative, the sets were examined, and outliers were removed using the interquartile range method because the distributions vary between the variables. The histogram, boxplot, and QQ plots were used to determine outliers that should be removed. Some variables, such as alcohol, total phenols, flavonoids, OD280, and nonflavonoid phenols did not have outliers. In most cases, the IQR method did not remove all of the outliers, but kept some upper and lower-bound outliers. Since the number of outliers was minimal, they were kept in the set as noise. Figure 5 is a sample of cleaned and uncleaned Ash Alcanity data. Please refer to the references section to see the cleaned version of each variable.



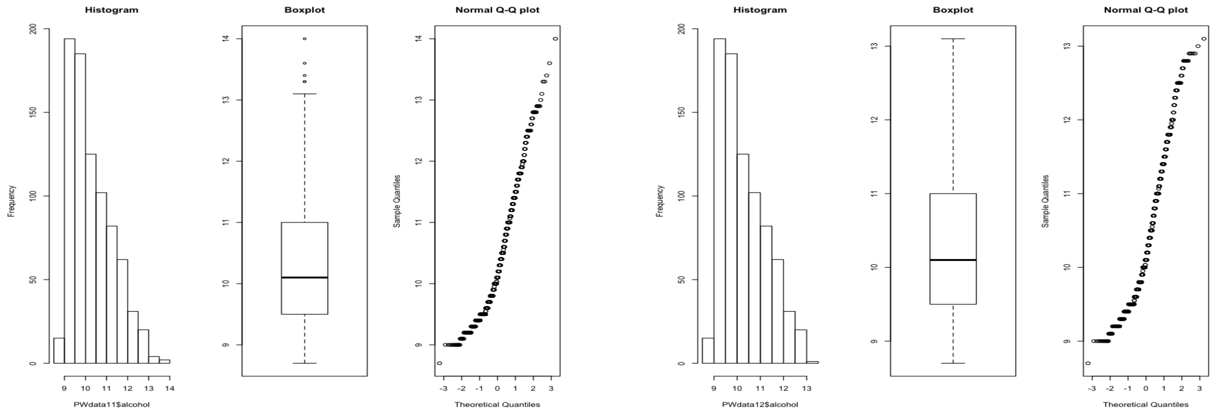
(a) Ash Alcanity Uncleaned Dataset

(b) Ash Alcanity Cleaned Dataset

Figure 5: Ash Alcanity Dataset

**Portugal Wine Dataset**

The Portugal Wine dataset is from the UCI Machine Learning repository. The dataset is related to red and white variants of the Portuguese Vinho Verde wine. Around 13 variables are present: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, alcohol, and quality. The other variables are considered when creating the discrete variable, quality. The Portugal Wine dataset is relatively large, with 1143 rows, and only contains quantitative data. All of the variables are complete and have no empty values. As a result, outliers were examined for cleaning the data. When cleaning this dataset, the free sulfur dioxide variable was removed from the set since total sulfur dioxide is present. The total sulfur dioxide is a combination of free sulfur dioxide and bounded sulfur dioxide. The same procedure was applied to examine outliers. Histograms, boxplots, and QQ plots were generated to examine outliers. After, the interquartile range method was utilized to eliminate outliers from the dataset since the distribution also varies. Figure 6 represents the uncleaned and cleaned graphs of the variables. Like in the other datasets, some outliers remained in the set as noise. Please review the references section for the dataset variables' uncleaned and cleaned plots. The cleaned dataset has around 804 rows which is slightly lower than the original uncleaned dataset.



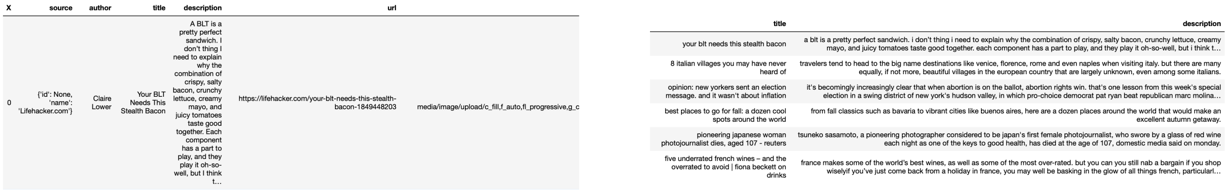
(a) Alcohol Uncleaned Dataset

(b) Alcohol Cleaned Dataset

Figure 6: Alcohol Dataset

**News API Text Dataset**

API data was gained through News API regarding current events on red wine. The dataset presented with 9 column variables. Those are X, source, author, title, description, url, urlToImage, publishedAt, and content. Based on examining the dataset, variables X, source, url, urlToImage, publishedAt, author, and content were removed from the dataset because they all contained insignificant information. The content variable had unreadable symbols, which was the basis for removing it from the dataset. Personal identifiers such as author were removed. The remaining variables are the title and description. News API dataset contains around 300 rows of data which is relatively large. Since the text data is from news resources, the assumption was made that no misspelled words were contained in the set. The dataset contained capital letters. As a result, all text data was converted to lowercase letters. Figure 7 is a comparison of uncleaned and cleaned data.



| X | source | author | title | description | url | |
|---|---|---|---|---|---|---|
| 0 | {'id': None, 'name': 'Lifehacker.com'} | Claire Lower | Your BLT Needs This Stealth Bacon | A BLT is a pretty perfect sandwich. I don't thing I need to explain why the combination of crispy, salty bacon, crunchy lettuce, creamy mayo, and juicy tomatoes taste good together. Each component has a part to play, and they play it oh-so-well, but I think t… | https://lifehacker.com/your-blt-needs-this-stealth-bacon-1849448203 | media/image/upload/c_fill,f_auto,fl_progressive,g_c |

(a) Uncleaned API Dataset

| title | description |
|---|---|
| your blt needs this stealth bacon | a blt is a pretty perfect sandwich. i don't thing i need to explain why the combination of crispy, salty bacon, crunchy lettuce, creamy mayo, and juicy tomatoes taste good together. each component has a part to play, and they play it oh-so-well, but i think t… |
| 8 italian villages you may have never heard of | travelers tend to head to the big name destinations like venice, florence, rome and even naples when visiting italy. but there are many equally, if not more, beautiful villages in the european country that are largely unknown, even among some italians. |
| opinion: new yorkers sent an election message. and it wasn't about inflation | it's becomingly increasingly clear that when abortion is on the ballot, abortion rights win. that's one lesson from this week's special election in a swing district of new york's hudson valley, in which pro-choice democrat pat ryan beat republican marc molina… |
| best places to go for fall: a dozen cool spots around the world | from fall classics such as bavaria to vibrant cities like buenos aires, here are a dozen places around the world that would make an excellent autumn getaway. |
| pioneering japanese woman photojournalist dies, aged 107 - reuters | tsuneko sasamoto, a pioneering photographer considered to be japan's first female photojournalist, who swore by a glass of red wine each night as one of the keys to good health, has died at the age of 107, domestic media said on monday. |
| five underrated french wines – and the overrated to avoid \| fiona beckett on drinks | france makes some of the world's best wines, as well as some of the most over-rated. but you can you still nab a bargain if you shop wiselyif you've just come back from a holiday in france, you may well be basking in the glow of all things french, particularl… |

(b) Cleaned API Dataset

Figure 7: APIDataset

## Results

Through machine learning, information about consumer satisfaction, the chemical makeup of good quality wine, and words associated with good wine can be explored. The Wine Enthusiast dataset can bring insight into the relationships between wine ratings and descriptions.
Certain words associated with high-rated wines can help characterize good quality wine and wine that consumers should seek. The Chemical Analysis dataset will allow the exploration of chemical compounds in wines and how they vary from bottle to bottle. A simple correlation plot between the variables in Chemical Analysis was performed. Figure 8 represents the variable correlation plot. The plot shows negative and positive correlations, which can help identify relationships between the compounds created during the production process. This information can be used to learn about the wine production process and optimization of congeners. The Portugal Wine dataset will help characterize what chemical compounds cause consumers to give high satisfaction ratings. Since this dataset provides the chemical composition of Portuguese wine with quality, the question of what chemical composition makes a good quality wine can be explored. Lastly, the API data will list words associated with wine in current news. Based on this data, the public attitude towards wine can be analyzed.

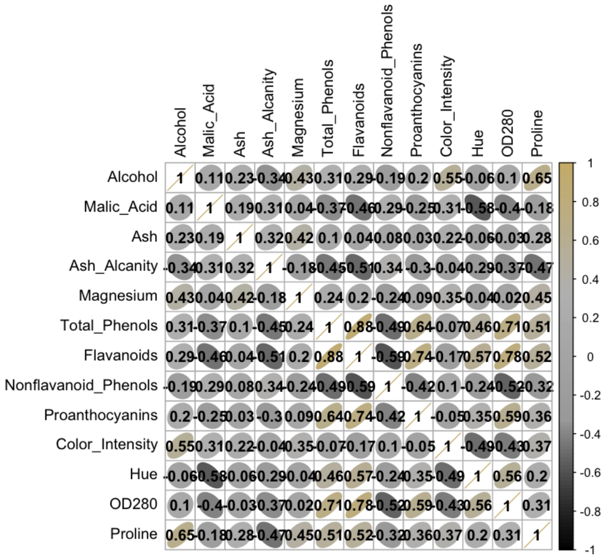


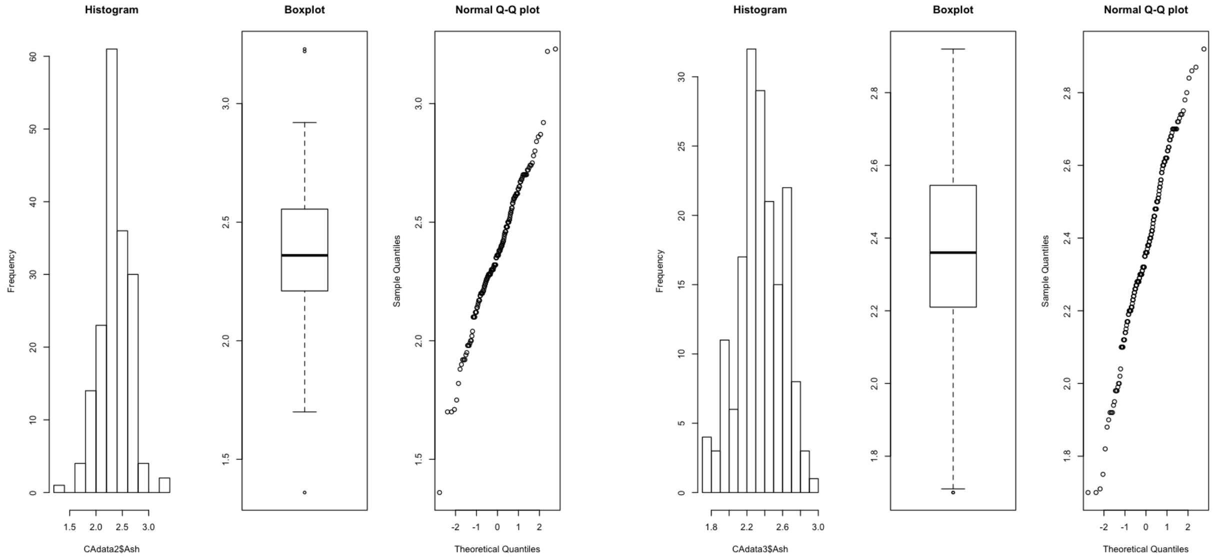
Figure 8: Chemical Analysis Correlation

## Conclusion

Through this project, relationships between the production process and wine quality will be explored. The goal of this project is to learn about how the production process influences wine quality and reactions from consumers. In this project, the focus will be on red wine. It would be interesting to explore the different properties of white wine in the future.

# References

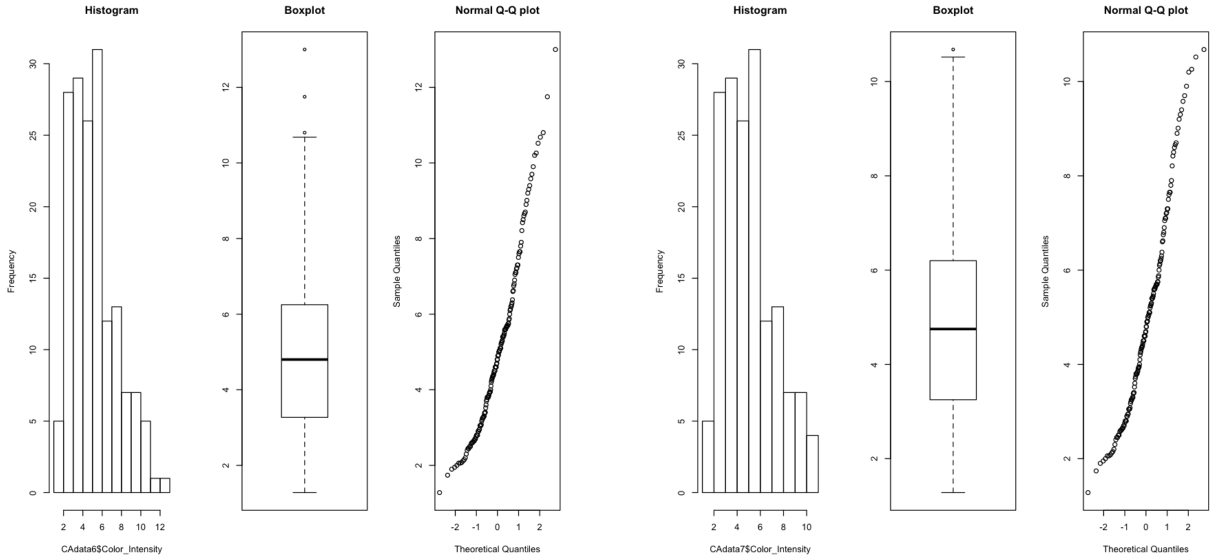
**Wine Enthusiast**
https://www.kaggle.com/datasets/zynicide/wine-reviews
**Chemical Analysis: Uncleaned and Cleaned Graphical Representation**
https://archive.ics.uci.edu/ml/datasets/wine



(a) Ash Uncleaned Dataset

(b) Ash Cleaned Dataset

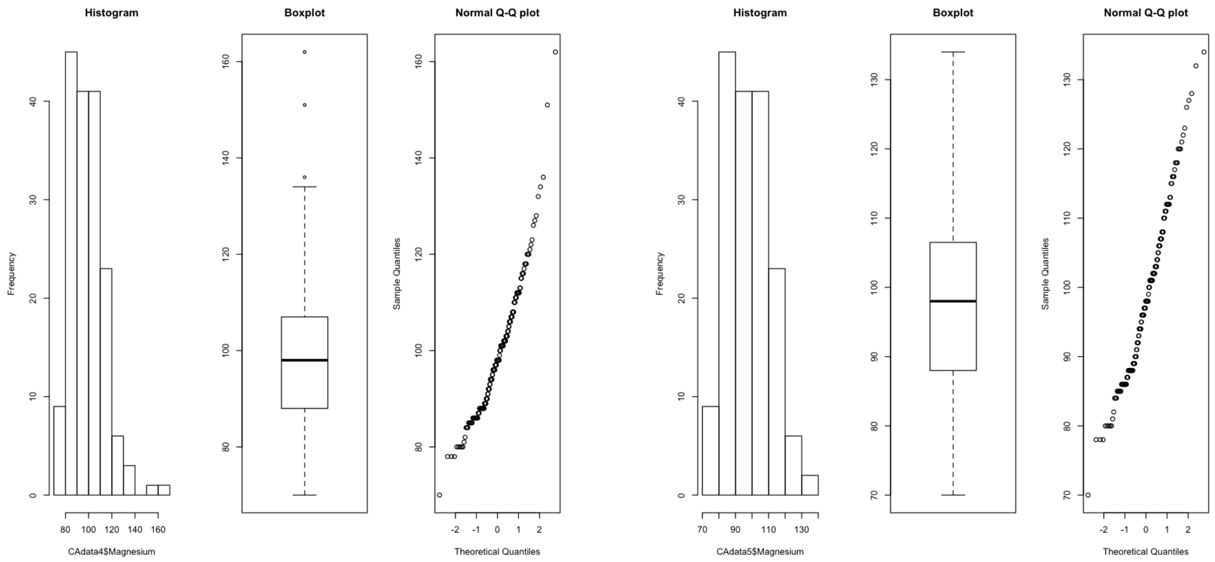
Figure 9: Ash Dataset



(a) Color Intensity Uncleaned Dataset

(b) Color Intensity Cleaned Dataset

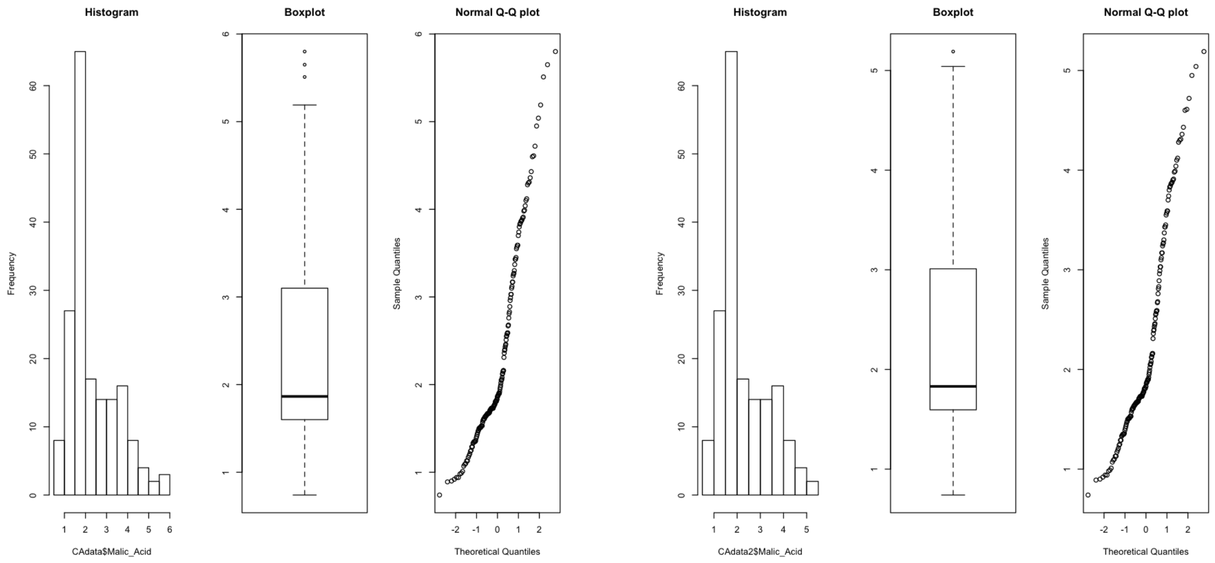
Figure 10: Color Intensity Dataset



(a) Magnesium Uncleaned Dataset

(b) Magnesium Cleaned Dataset

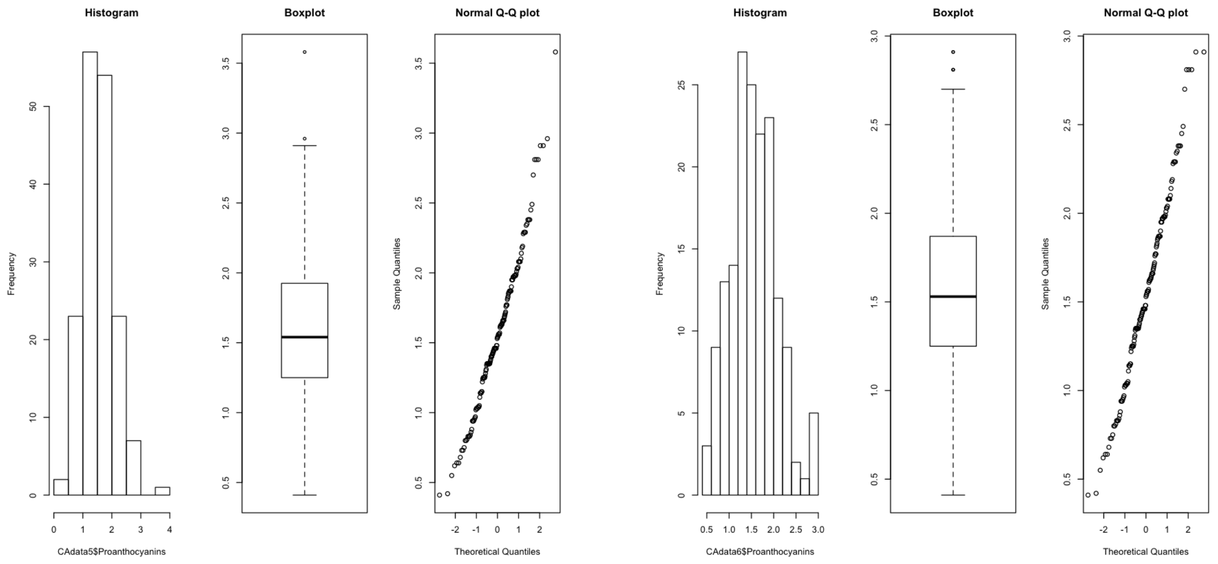
Figure 11: Magnesium Dataset



(a) Malic Acid Uncleaned Dataset

(b) Malic Acid Cleaned Dataset

Figure 12: Malic Acid Dataset



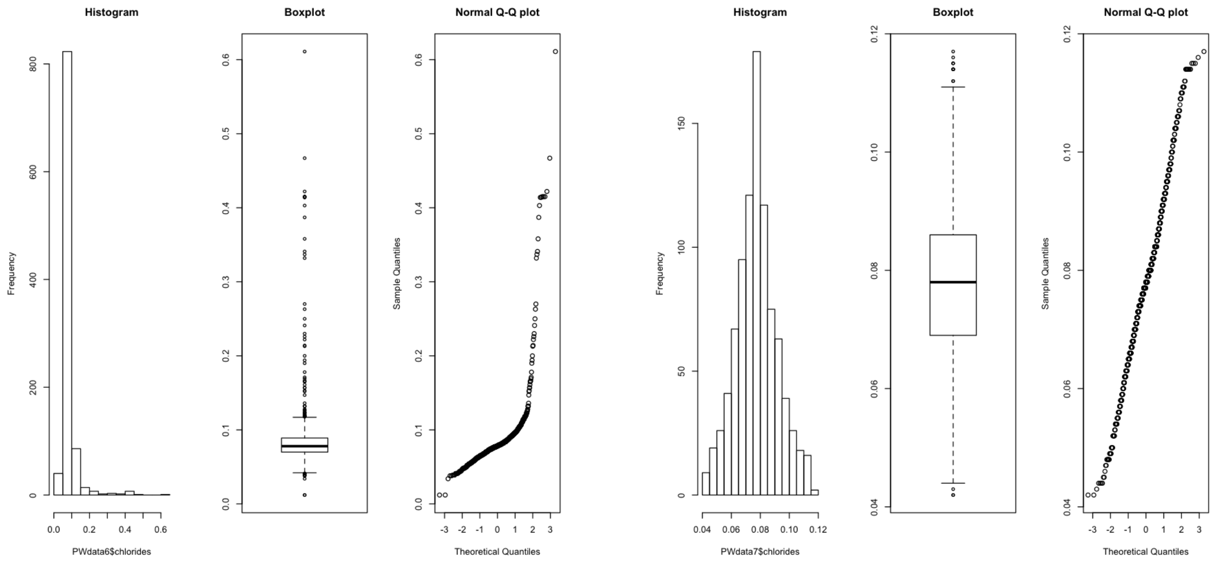
(a) Proanthocyanins Uncleaned Dataset

(b) Proanthocyanins Cleaned Dataset

Figure 13: Proanthocyanins Dataset

## Portugal Wine: Uncleaned and Cleaned Graphical Representation

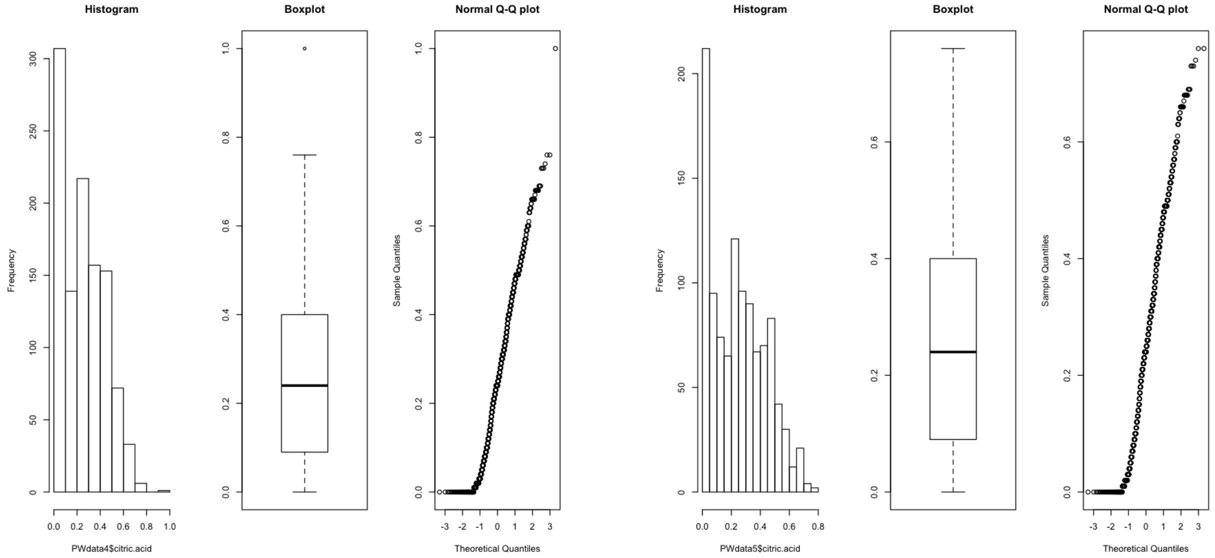
https://archive.ics.uci.edu/ml/datasets/wine+quality



(a) Chlorides Uncleaned Dataset

(b) Chlorides Cleaned Dataset

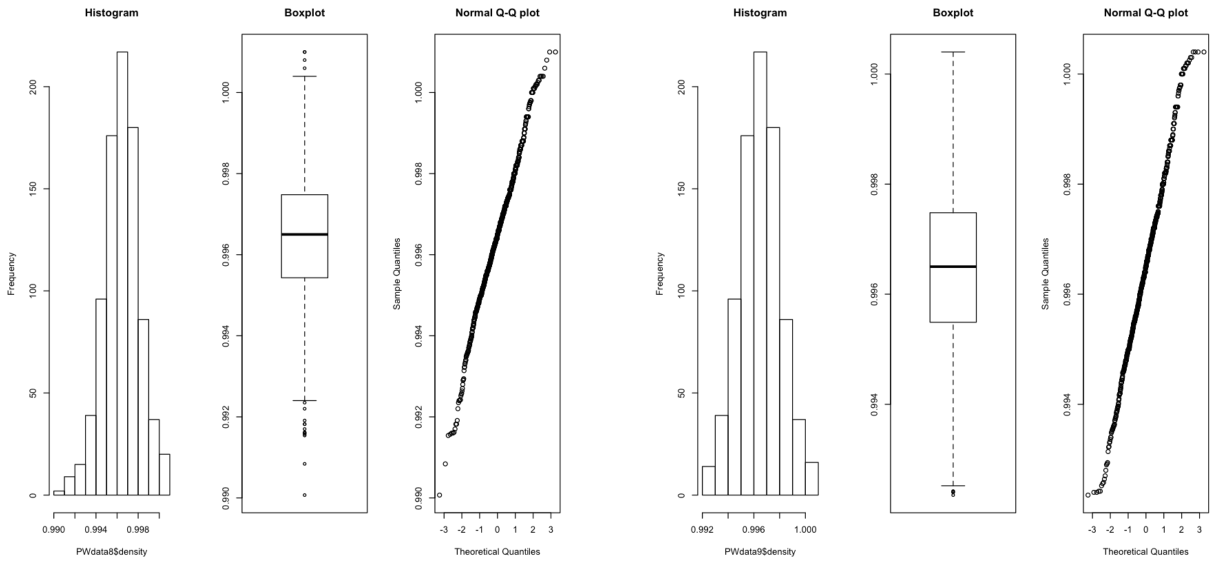
Figure 14: Chlorides Dataset



(a) Citric Acid Uncleaned Dataset

(b) Citric Acid Cleaned Dataset

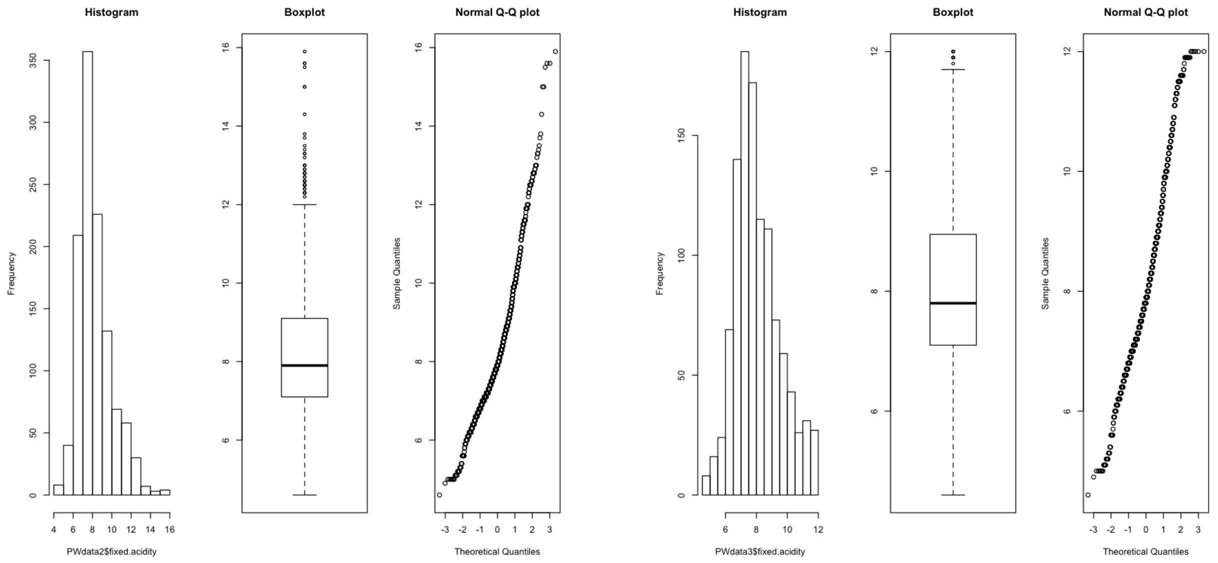
Figure 15: Citric Acid Dataset



(a) Density Dataset

(b) Density Dataset

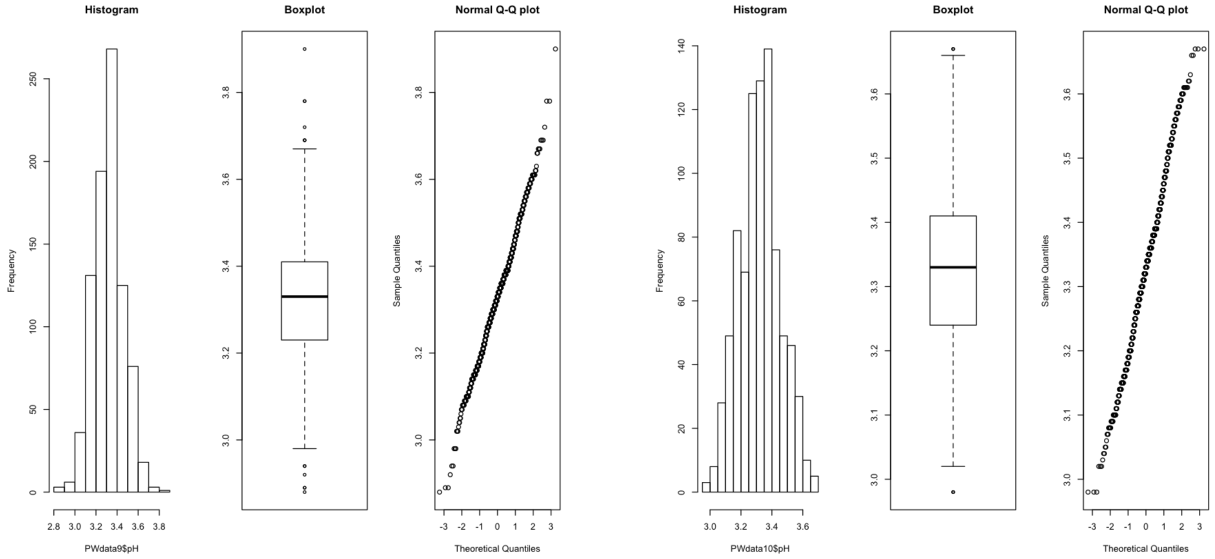
Figure 16: Density Dataset



(a) Fixed Acidity Uncleaned Dataset

(b) Fixed Acidity Cleaned Dataset

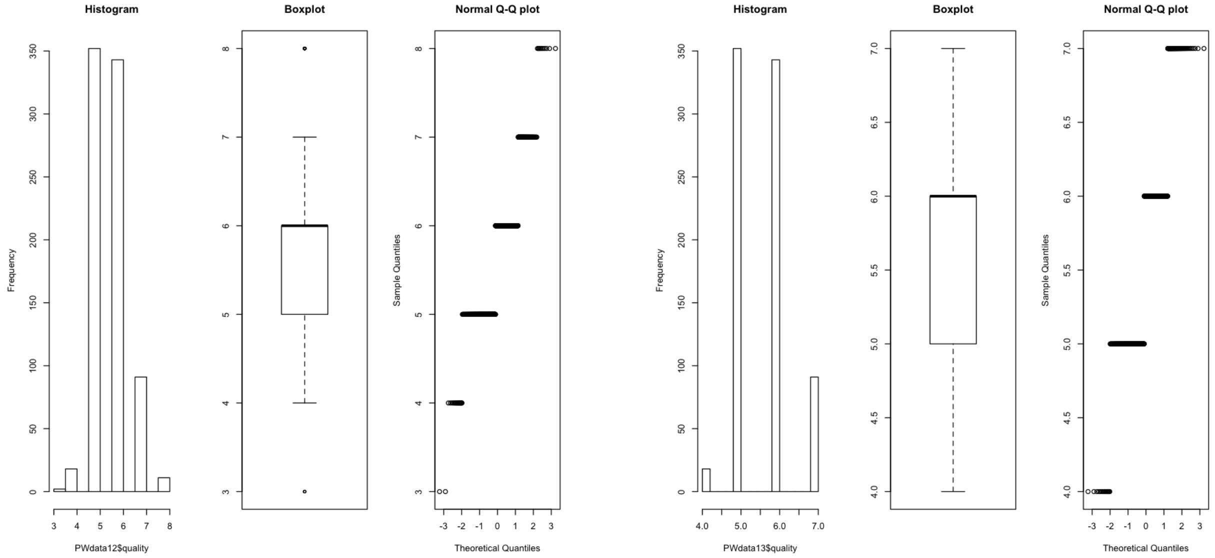
Figure 17: Fixed Acidity Dataset



(a) pHUncleaned Dataset

(b) pH Cleaned Dataset

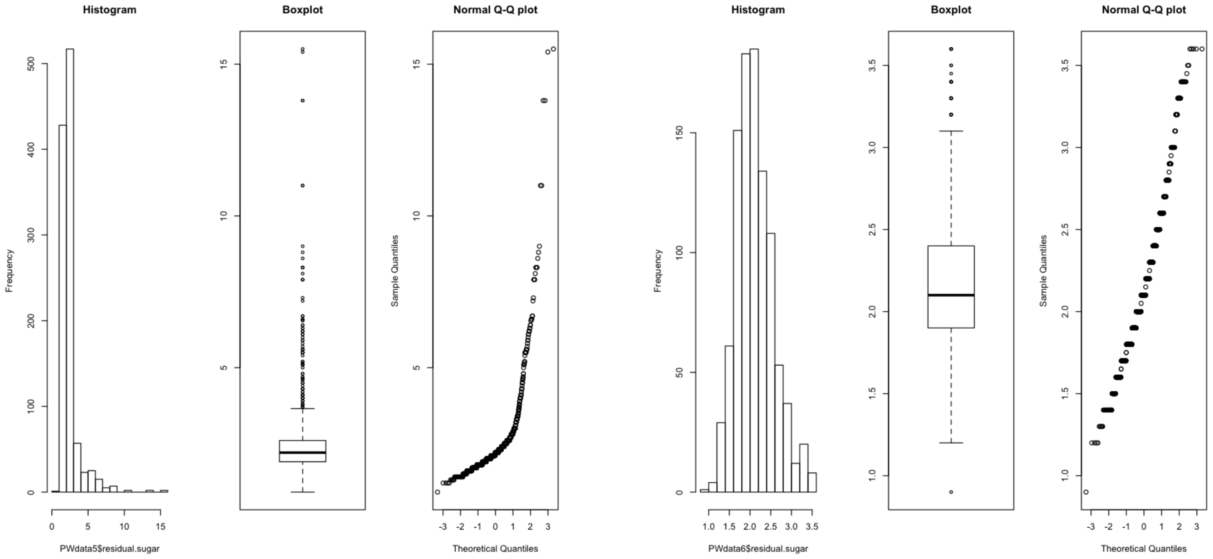
Figure 18: pH Dataset



(a) Quality Uncleaned Dataset

(b) Quality Cleaned Dataset

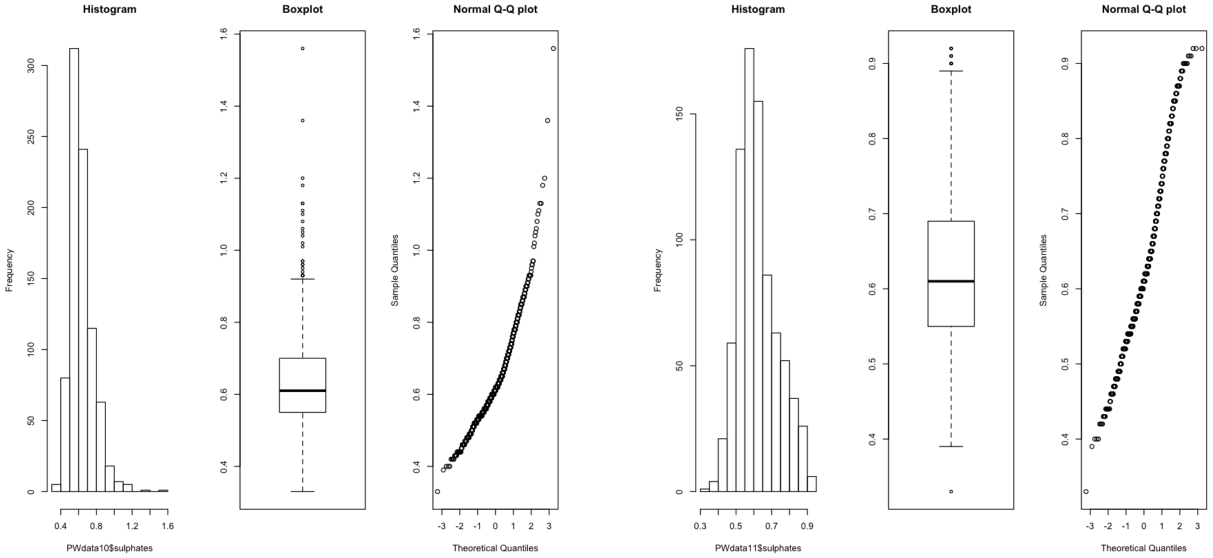
Figure 19: Quality Dataset



(a) Residual Sugars Uncleaned Dataset

(b) Residual Sugars Cleaned Dataset

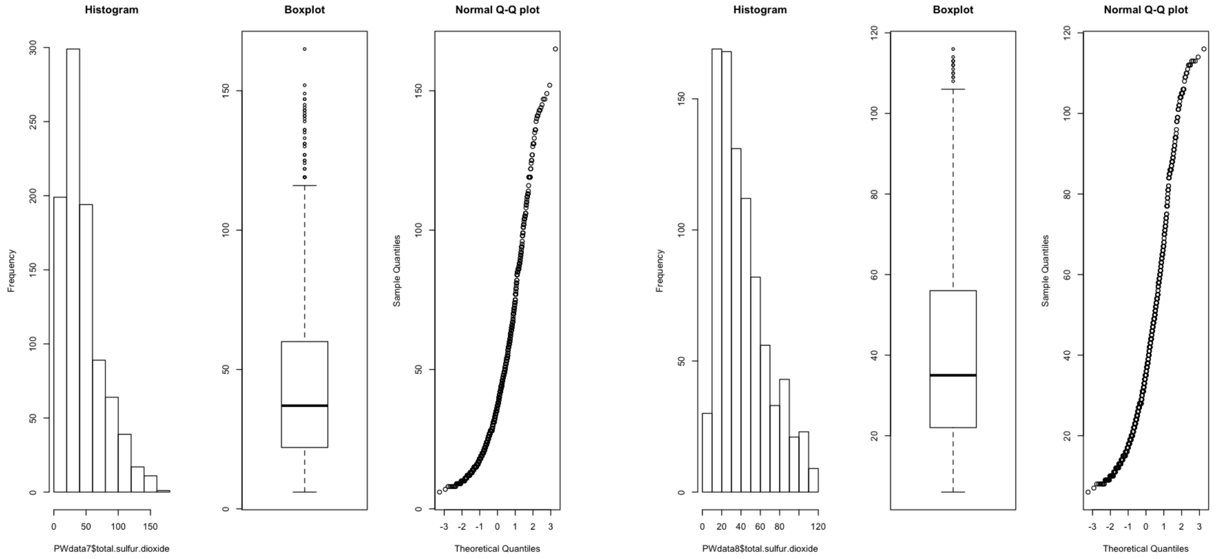
Figure 20: Residual Sugars Dataset



(a) Sulphates Uncleaned Dataset

(b) Sulphates Cleaned Dataset

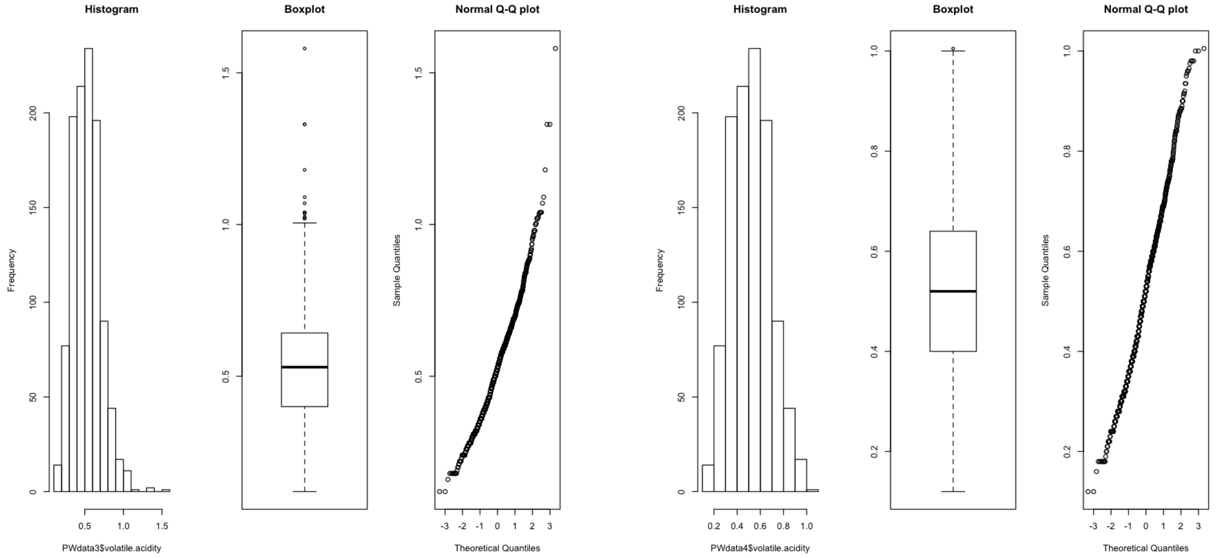
Figure 21: Sulphates Dataset



(a) Total Sulfur Dioxide Uncleaned Dataset

(b) Total Sulfur Dioxide Dataset

Figure 22: Total Sulfur Dioxide Dataset



(a) Volatile Acidity Uncleaned Dataset

(b) Volatile Acidity Cleaned Dataset

Figure 23: Volatile Acidity Dataset