WINE QUALITY

Final Project Report

Dec 10, 2015

Prepared by: Abhinav Singh

Stevens Institute of Technology

**Table of Content**

1 CRISP DM ………………………………………………………………………………………… 3

2 Business Understanding ………………………………………………………………….. 6

3 Data Understanding - Dataset …………………………………………………………. 7

4 Data Understanding – Exploratory Analysis ...…………………………………. 9

5 Principal Component Analysis ………………………………………………………… 17

6 Factor Analysis ………………………………………………………………………………. 18

7 Data Preparation …………………………………………………………………………… 19

8 Modeling – Red Wine ………………………………………………………………….… 20

9 Modeling – White Wine ……………………………………………….……………….. 21

10 Modeling – All Dataset……………………………………………………..………….. 22

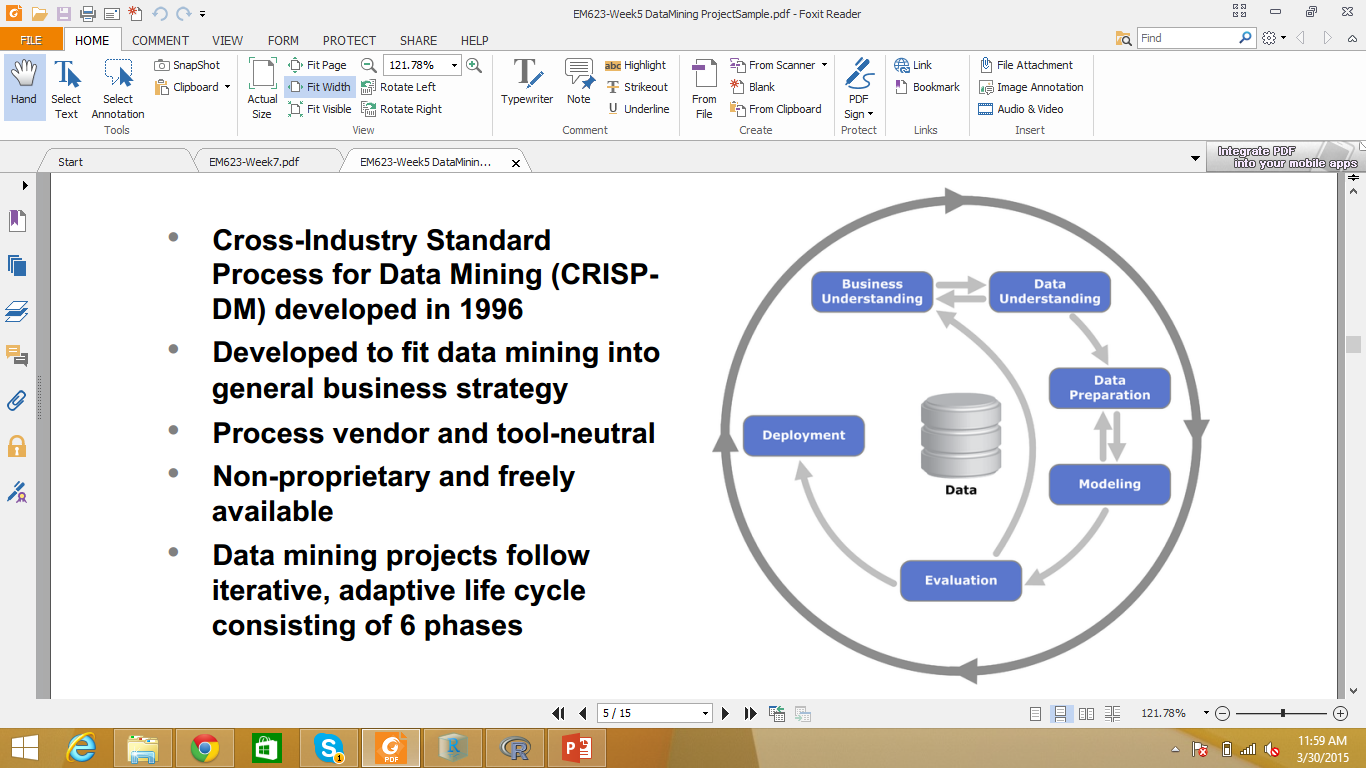
11 Evaluation – White Wine …………………………………………………………….. 23

12 Evaluation – All Wine ………………………………………………………………….. 24

Appendix: SAS program code …………………………………………………………… 25

**The CRISP-DM methodology**

* We will use the CRISP (Cross-Industry Standard Process for Data Mining) methodology to give the directions of this project. This process, consists in a iterative and adaptive life cycle consisting of 6 phases, as shown on the below picture.



The life cycle of a data mining project consists of six phases, as shown in the above figure. The sequence of the phases is not rigid. Moving back and forth between different phases is always required. 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. The outer circle in the above figure symbolizes the cyclical nature of data mining itself. Data mining does not end once a solution is deployed. The lessons learned during the process and from the deployed solution can trigger new, often more-focused business questions. Subsequent data mining processes will benefit from the experiences of previous ones. In the following, we briefly outline each phase:

**Business understanding**

This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

**Data understanding**

The data understanding phase starts with initial data collection and proceeds with activities that will enable us to become familiar with the data, identify data quality problems, discover first insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information.

**Data preparation**

The data preparation phase covers all activities needed to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools.

**Modeling**

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, going back to the data preparation phase is often necessary.

**Evaluation**

At this stage in the project, we have built a model (or models) that appears to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to thoroughly evaluate it and review the steps executed to create it, to be certain the model properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached. Deployment Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. It often involves applying “live” models within an organization’s decision making processes—for example, real-time personalization of Web pages or repeated scoring of marketing databases. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. In many cases, it is the customer, not the data analyst, who carries out the deployment steps. However, even if the analyst will carry out the deployment effort, it is important for the customer to understand up front what actions need to be carried out in order to actually make use of the created models.

**Business Understanding**

Once viewed as a luxury good, nowadays wine is increasingly enjoyed by a wider range of consumers. Portugal is a top ten wine exporting country with 3.17% of the market share in 2005. Exports of its vinho verde wine (from the northwest region) have increased by 36% from 1997 to 2007. To support its growth, the wine industry is investing in new technologies for both wine making and selling processes. Wine certification and quality assessment are key elements within this context. Certification prevents the illegal adulteration of wines (to safeguard human health) and assures quality for the wine market. Quality evaluation is often part of the certification process and can be used to improve wine making (by identifying the most influential factors) and to stratify wines such as premium brands (useful for setting prices).

Wine certification is generally assessed by physicochemical and sensory tests. Physicochemical laboratory tests routinely used to characterize wine include determination of density, alcohol or pH values, while sensory tests rely mainly on human experts

This study will consider vinho verde, a unique product from the Minho (northwest) region of Portugal. Medium in alcohol, is it particularly appreciated due to its freshness (specially in the summer). This wine accounts for 15% of the total Portuguese production and around 10% is exported, mostly white wine. In this work, we will analyze the two most common variants, white and red (ros´e is also produced), from the demarcated region of vinho verde. The data were collected from May/2004 to February/2007.

**Data Understanding – Dataset**

We have taken the following three datasets:-

1. Red Wine dataset with 1599 observations
2. White Wine dataset with 4898 observations
3. All dataset which is a combination of Red Wine dataset and White Wine dataset and hence consists of 6497 (1599+4898) observations.

**Variables:**

Fixed Acidity

Volatile Acidity

Citric Acid

Residual Sugar

Chlorides

Free sulfur dioxide

total sulfur dioxide

density

pH

sulphates

alcohol

quality

**Description of Variables**

1 - fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

2 - volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

3 - citric acid: found in small quantities, citric acid can add ‘freshness’ and flavor to wines

4 - residual sugar: the amount of sugar remaining after fermentation stops, it’s rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet

5 - chlorides: the amount of salt in the wine

6 - free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

7 - total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine

8 - density: the density of water is close to that of water depending on the percent alcohol and sugar content

9 - pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale

10 - sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant

11 - alcohol: the percent alcohol content of the wine

**Data Understanding - Exploratory Analysis**

In order to understand the data, we performed exploratory analysis on every variable. The result of the exploratory analysis is given below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Fixed Acidity** | | | |  |  |  |
|  | **All** | **Red** | **White** | **Red-White** | **Ratio** | **ratio 2** |
| **Number of observations** | **6497** | **1599** | **4898** |  |  |  |
| **Mean** | **7.215** | **8.31** | **6.854** | 1.456 | 17.52% | 21.2% |
| **Std Deviation** | **1.296** | **1.74** | **0.843** | 0.897 | 51.55% | 106.4% |
| **Variance** | **1.68** | **3.03** | **0.712** | 2.318 | 76.50% | 325.6% |
| **Median** | **7** | **7.9** | **6.8** | 1.1 | 13.92% | 16.2% |
| **Mode** | **6.8** | **7.2** | **6.8** | 0.4 | 5.56% | 5.9% |
| **Highest** | **15.9** | **15.9** | **14.2** | 1.7 | 10.69% | 12.0% |
| **Lowest** | **3.8** | **4.6** | **3.8** | 0.8 | 17.39% | 21.1% |

Based on the exploratory analysis of Fixed Acidity, we can say that in average, the white wine is 17.52 more acidic than the red wine.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Volatile Acidity** | | | |  |  |  |
|  | **All** | **red** | **White** |  |  |  |
| **Number of observations** | **6497** | **1599** | **4898** | -3299 |  | -67.4% |
| **Mean** | **0.339** | **0.52** | **0.278** | 0.242 | 46.5% | 87.1% |
| **Std Deviation** | **0.164** | **0.17** | **0.1** | 0.07 | 41.2% | 70.0% |
| **Variance** | **0.027** | **0.034** | **0.01** | 0.024 | 70.6% | 240.0% |
| **Median** | **0.29** | **0.52** | **0.26** | 0.26 | 50.0% | 100.0% |
| **Mode** | **0.28** | **0.6** | **0.28** | 0.32 | 53.3% | 114.3% |
| **Highest** | **1.58** | **1.58** | **1.1** | 0.48 | 30.4% | 43.6% |
| **Lowest** | **0.08** | **0.12** | **0.08** | 0.04 | 33.3% | 50.0% |

Based on the exploratory analysis of Volatile Acidity, we can say that in average, the volatile acidity for red wine is 87.1% higher than that of the white wine.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Citric Acid** | | | |  |  |  |
|  | **All** | **red** | **White** |  |  |  |
| **Number of observations** | **6497** | **1599** | **4898** | -3299 | -206.3% | -67.4% |
| **Mean** | **0.318** | **0.271** | **0.334** | -0.063 | -23.2% | -18.9% |
| **Std Deviation** | **0.145** | **0.195** | **0.121** | 0.074 | 37.9% | 61.2% |
| **Variance** | **0.021** | **0.037** | **0.014** | 0.023 | 62.2% | 164.3% |
| **Median** | **0.31** | **0.26** | **0.32** | -0.06 | -23.1% | -18.8% |
| **Mode** | **0.3** | **0** | **0.3** | -0.3 | #DIV/0! | -100.0% |
| **Highest** | **1.66** | **1** | **1.66** | -0.66 | -66.0% | -39.8% |
| **Lowest** | **0** | **0** | **0** | 0 | #DIV/0! | #DIV/0! |

Based on the exploratory analysis of Citric Acid, we can say that in average, the white wine has 23.2% more critic acid than the red wine.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Residual Sugar** | | | |  |  |  |
|  | **All** | **red** | **White** |  |  |  |
| **Number of observations** | **6497** | **1599** | **4898** | -3299 | -206.3% | -67.4% |
| **Mean** | **5.443** | **2.538** | **6.391** | -3.853 | -151.8% | -60.3% |
| **Std Deviation** | **4.757** | **1.409** | **5.072** | -3.663 | -260.0% | -72.2% |
| **Variance** | **22.636** | **1.987** | **25.72** | -23.733 | -1194.4% | -92.3% |
| **Median** | **3** | **2.2** | **5.2** | -3 | -136.4% | -57.7% |
| **Mode** | **2** | **2** | **1.2** | 0.8 | 40.0% | 66.7% |
| **Highest** | **65.8** | **15.5** | **65.8** | -50.3 | -324.5% | -76.4% |
| **Lowest** | **0.6** | **0.9** | **0.6** | 0.3 | 33.3% | 50.0% |

Based on the exploratory analysis of Residual Sugar, we can say that in average the white wine has 151.8% more sugar than the red wine

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Chlorides** | | | |  |  |  |
|  | **All** | **red** | **White** |  |  |  |
| **Number of observations** | **6497** | **1599** | **4898** | -3299 | -206.3% | -67.4% |
| **Mean** | **0.056** | **0.087** | **0.045** | 0.042 | 48.3% | 93.3% |
| **Std Deviation** | **0.035** | **0.047** | **0.021** | 0.026 | 55.3% | 123.8% |
| **Variance** | **0.001** | **0.002** | **0** | 0.002 | 100.0% | #DIV/0! |
| **Median** | **0.047** | **0.079** | **0.043** | 0.036 | 45.6% | 83.7% |
| **Mode** | **0.044** | **0.08** | **0.044** | 0.036 | 45.0% | 81.8% |
| **Highest** | **0.611** | **0.611** | **0.346** | 0.265 | 43.4% | 76.6% |
| **Lowest** | **0.009** | **0.012** | **0.009** | 0.003 | 25.0% | 33.3% |

Based on the exploratory analysis of Chlorides, we can say that in average the red wine has 93.3% more chlorides than the white wine.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **free\_sulfur\_dioxide** | | | |  |  |  |
|  | **All** | **red** | **White** |  |  |  |
| **Number of observations** | **6497** | **1599** | **4898** | -3299 | -206.3% | -67.4% |
| **Mean** | **30.525** | **15.874** | **35.308** | -19.434 | -122.4% | -55.0% |
| **Std Deviation** | **17.749** | **10.46** | **17.007** | -6.547 | -62.6% | -38.5% |
| **Variance** | **315.04** | **109.41** | **289.242** | -179.832 | -164.4% | -62.2% |
| **Median** | **29** | **14** | **34** | -20 | -142.9% | -58.8% |
| **Mode** | **29** | **6** | **29** | -23 | -383.3% | -79.3% |
| **Highest** | **289** | **72** | **289** | -217 | -301.4% | -75.1% |
| **Lowest** | **1** | **1** | **2** | -1 | -100.0% | -50.0% |

Based on the exploratory analysis of free\_sulfur\_dioxide, we can say that in average the white wine has 122.4% more free sulfur dioxide than the red wine.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **total\_sulfur\_dioxide** | | | |  |  |  |
|  | **All** | **red** | **White** |  |  |  |
| **Number of observations** | **6497** | **1599** | **4898** | -3299 | -206.3% | -67.4% |
| **Mean** | **115.74** | **46.467** | **138.36** | -91.893 | -197.8% | -66.4% |
| **Std Deviation** | **56.52** | **32.895** | **42.498** | -9.603 | -29.2% | -22.6% |
| **Variance** | **3194.72** | **1082.1** | **1806.08** | -723.98 | -66.9% | -40.1% |
| **Median** | **118** | **38** | **134** | -96 | -252.6% | -71.6% |
| **Mode** | **111** | **28** | **111** | -83 | -296.4% | -74.8% |
| **Highest** | **440** | **289** | **440** | -151 | -52.2% | -34.3% |
| **Lowest** | **6** | **6** | **9** | -3 | -50.0% | -33.3% |

Based on the exploratory analysis of total\_sulfur\_dioxide, we can say that the white wine has 197.8% more total sulfur dioxide than the red wine.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **pH** | | | |  |  |  |  |
|  | **All** | **red** | **White** |  |  |  |  |
| **Number of observations** | **6497** | **1599** | **4898** | -3299 | -206.3% | -67.4% |  |
| **Mean** | **3.218** | **3.31** | **3.188** | 0.122 | 3.7% | 3.8% |  |
| **Std Deviation** | **0.16** | **0.154** | **0.151** | 0.003 | 1.9% | 2.0% |  |
| **Variance** | **0.025** | **0.023** | **0.022** | 0.001 | 4.3% | 4.5% |  |
| **Median** | **3.21** | **3.31** | **3.18** | 0.13 | 3.9% | 4.1% |  |
| **Mode** | **3.16** | **3.3** | **3.14** | 0.16 | 4.8% | 5.1% |  |
| **Highest** | **4.01** | **4.01** | **3.82** | 0.19 | 4.7% | 5.0% |  |
| **Lowest** | **2.72** | **2.74** | **2.72** | 0.02 | 0.7% | 0.7% |  |

Based on the exploratory analysis of pH, we can say that

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sulphates** | | | |  |  |  |
|  | **All** | **red** | **White** |  |  |  |
| **Number of observations** | **6497** | **1599** | **4898** | -3299 | -206.3% | -67.4% |
| **Mean** | **0.531** | **0.658** | **0.489** | 0.169 | 25.7% | 34.6% |
| **Std Deviation** | **0.148** | **0.169** | **0.114** | 0.055 | 32.5% | 48.2% |
| **Variance** | **0.022** | **0.02** | **0.013** | 0.007 | 35.0% | 53.8% |
| **Median** | **0.51** | **0.62** | **0.47** | 0.15 | 24.2% | 31.9% |
| **Mode** | **0.5** | **0.6** | **0.5** | 0.1 | 16.7% | 20.0% |
| **Highest** | **2** | **2** | **1.08** | 0.92 | 46.0% | 85.2% |
| **Lowest** | **0.22** | **0.33** | **0.22** | 0.11 | 33.3% | 50.0% |

Based on the exploratory analysis of Sulphates, we can say that we have 34.6% more sulphates in the red wine than the white wine.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Alcohol** | | | |  |  |  |
|  | **All** | **red** | **White** |  |  |  |
| **Number of observations** | **6497** | **1599** | **4898** | -3299 | -206.3% | -67.4% |
| **Mean** | **10.491** | **10.422** | **10.514** | -0.092 | -0.9% | -0.9% |
| **Std Deviation** | **1.192** | **1.065** | **1.23** | -0.165 | -15.5% | -13.4% |
| **Variance** | **1.422** | **1.134** | **1.514** | -0.38 | -33.5% | -25.1% |
| **Median** | **10.3** | **10.2** | **10.4** | -0.2 | -2.0% | -1.9% |
| **Mode** | **9.5** | **9.5** | **9.4** | 0.1 | 1.1% | 1.1% |
| **Highest** | **14.9** | **14.9** | **14.2** | 0.7 | 4.7% | 4.9% |
| **Lowest** | **8** | **8.4** | **8** | 0.4 | 4.8% | 5.0% |

Based on the exploratory analysis of Alcohol, we can say that

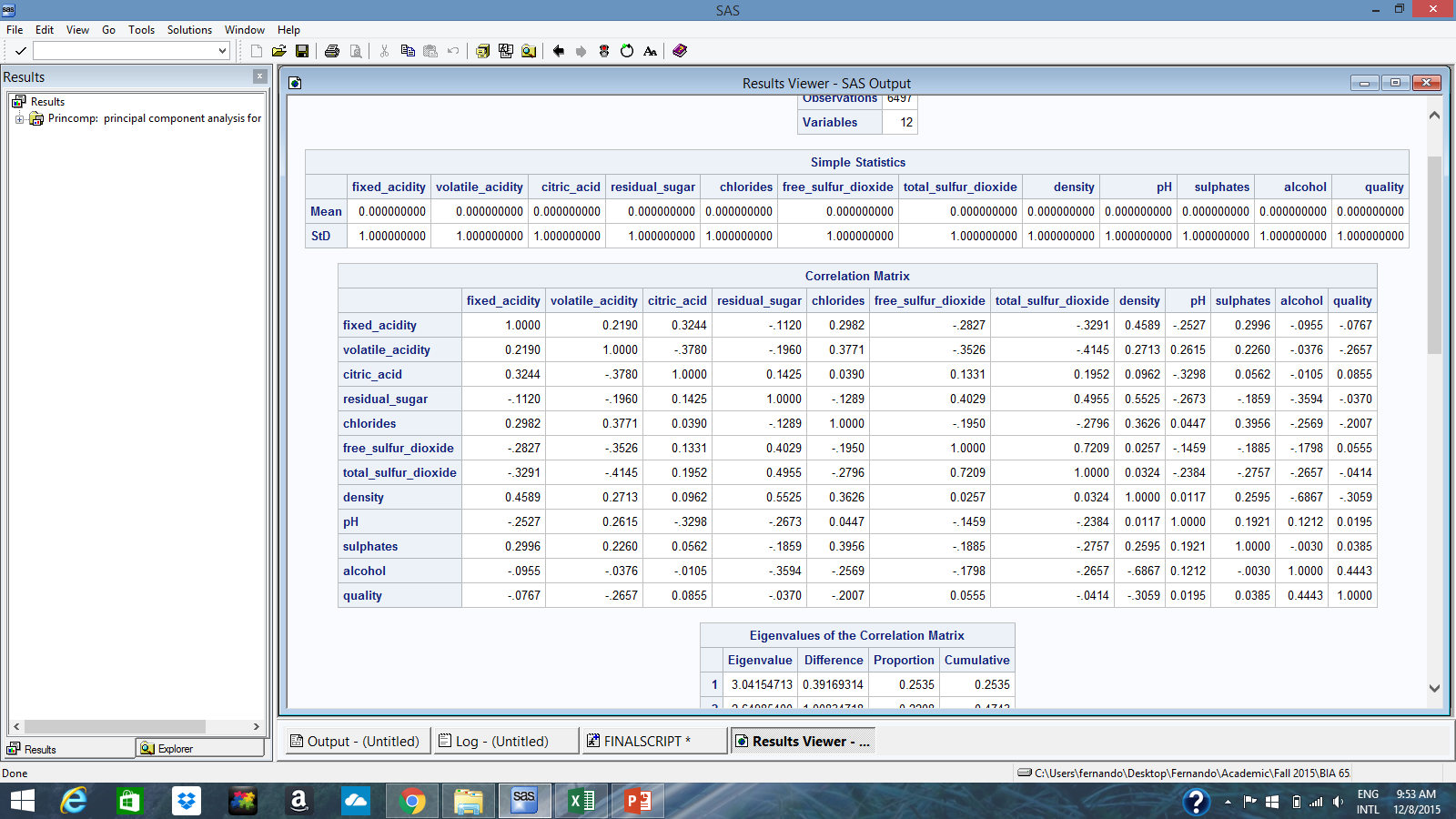
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Quality** | | | |  |  |  |
|  | **All** | **red** | **White** |  |  |  |
| **Number of observations** | **6497** | **1599** | **4898** | -3299 | -206.3% | -67.4% |
| **Mean** | **5.818** | **5.63** | **5.877** | -0.247 | -4.4% | -4.2% |
| **Std Deviation** | **0.873** | **0.807** | **0.886** | -0.079 | -9.8% | -8.9% |
| **Variance** | **0.762** | **0.652** | **0.784** | -0.132 | -20.2% | -16.8% |
| **Median** | **6** | **6** | **6** | 0 | 0.0% | 0.0% |
| **Mode** | **6** | **5** | **6** | -1 | -20.0% | -16.7% |
| **Highest** | **9** | **8** | **9** | -1 | -12.5% | -11.1% |
| **Lowest** | **3** | **3** | **3** | 0 | 0.0% | 0.0% |

Based on the exploratory analysis of Quality, we can say that

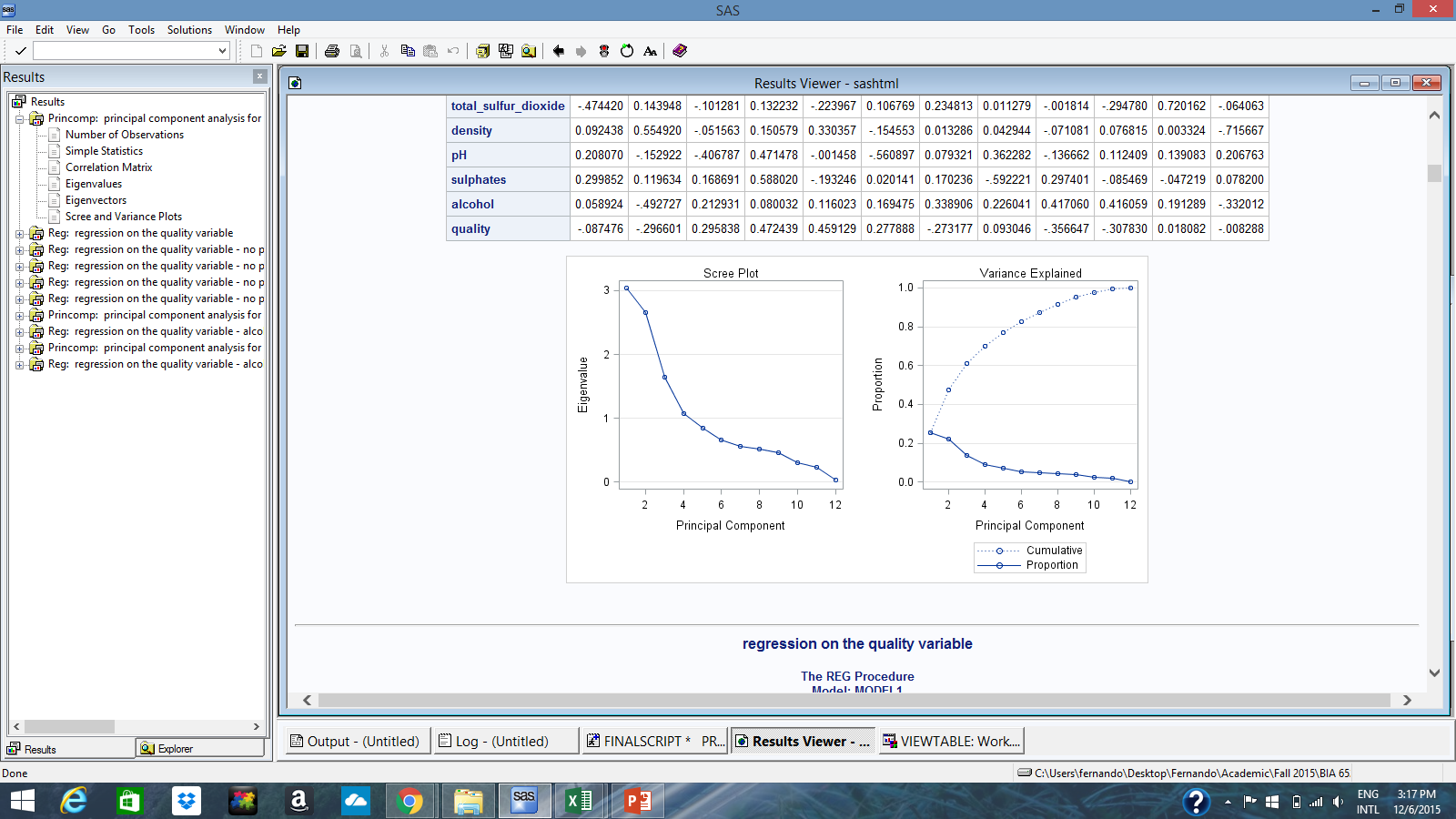
**Principal Component Analysis**

**All dataset**

Performing Principal Component Analysis on all dataset, we got the following result:

****

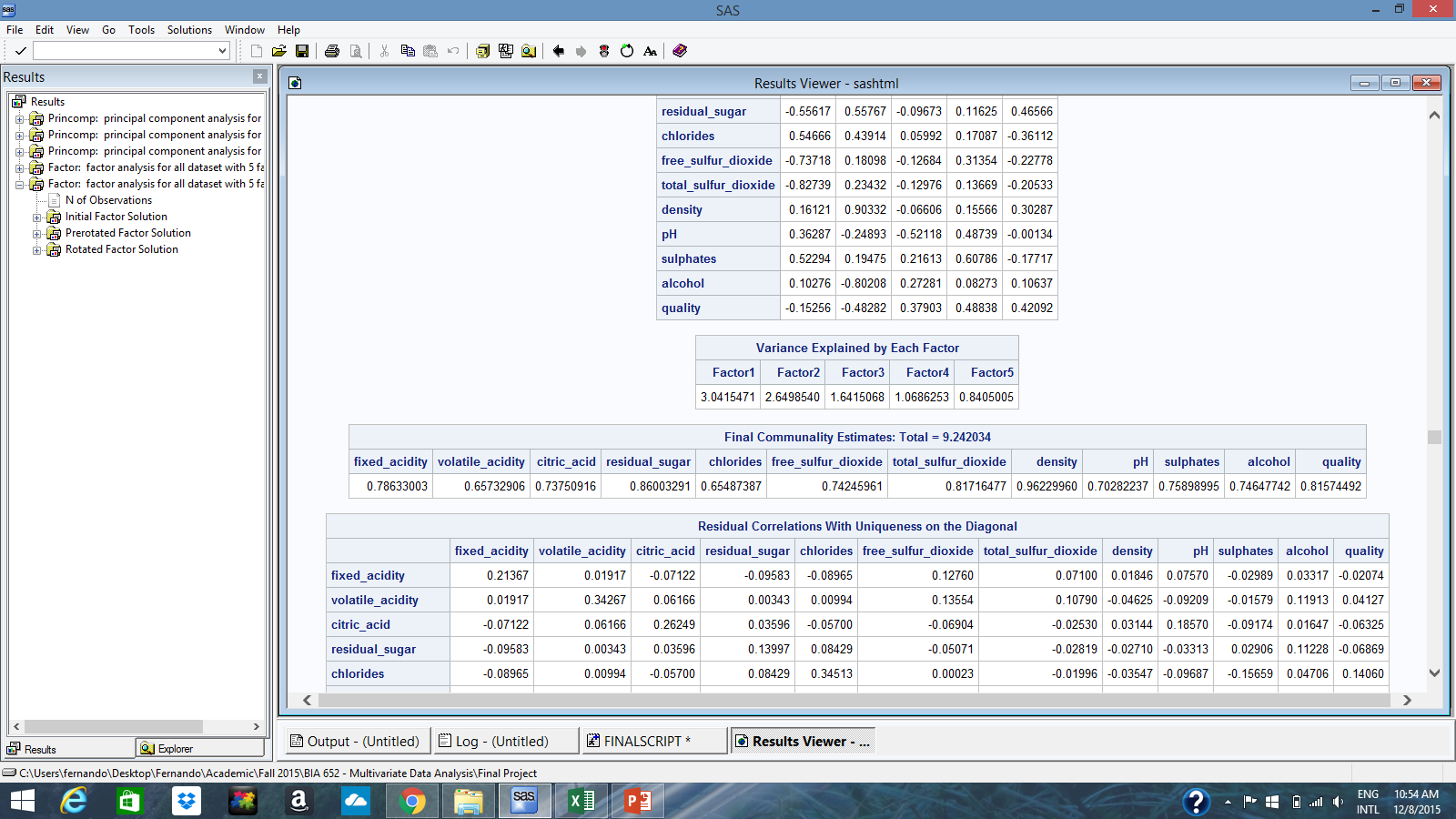
Analysis: We can clearly observe that the data is not correlated.



Analysis: We can clearly see that the same pattern repeats to white and red

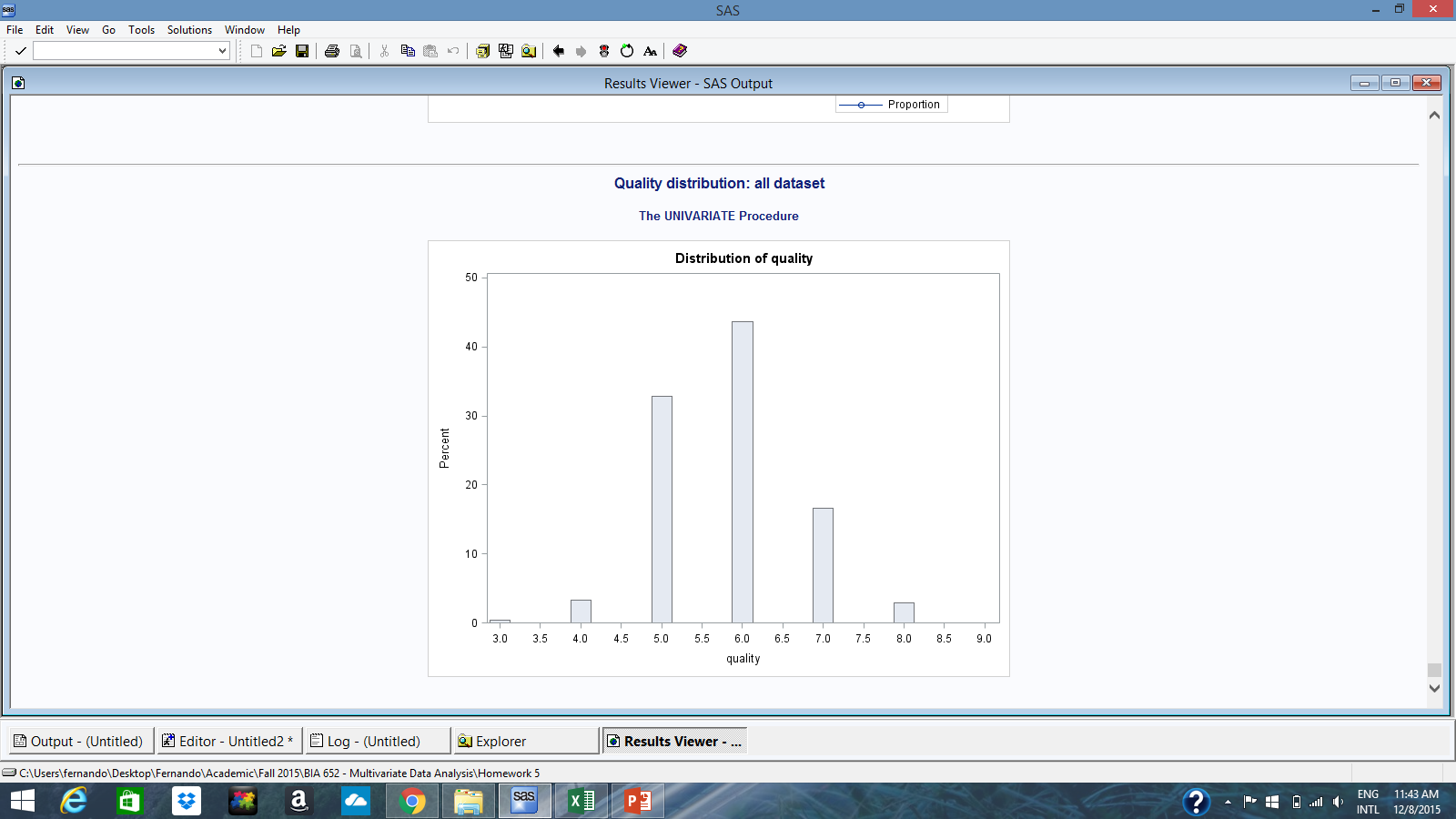
**Factor Analysis**

On performing factor analysis of 5 factors, we got the following output:



Analysis: We can clearly observe that Factor Analysis is also not a good method.

**Data Preperation**



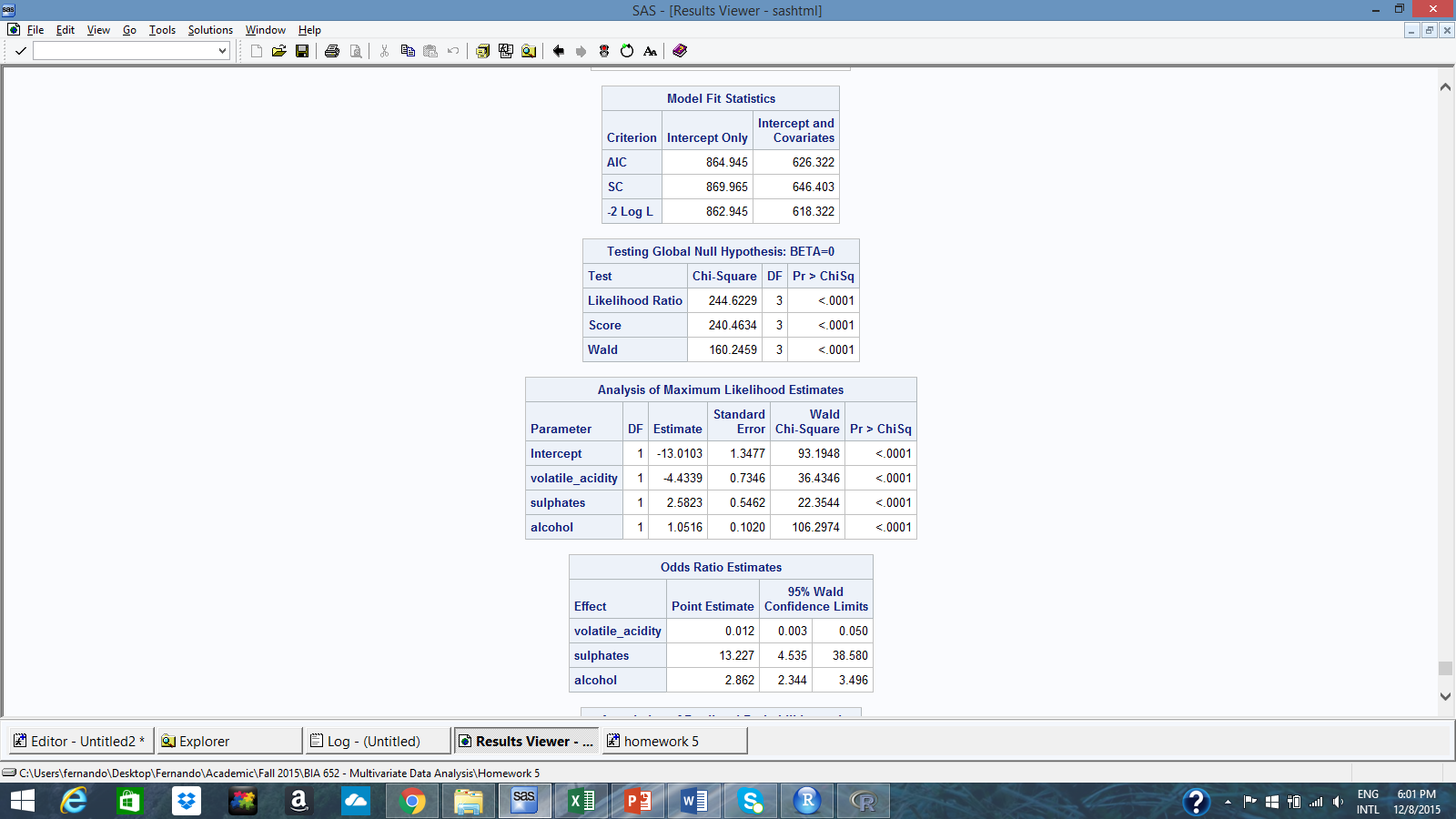
* There are no missing values
* We are creating a model to predict quality
* We have performed the following data modification:
* *If wine quality is higher than ‘6’ then we are assigning it a binary value of ‘1’*
* *If wine quality is less than or equal to ‘6’ then we are assigning it a binary value of ‘0’*

Analysis: We observed that only 20% of our data has high quality after modification!

**Modeling – Red Wine**

In order to model the red wine, we performed logistic regression on the red wine dataset. The following variables gave the best model:

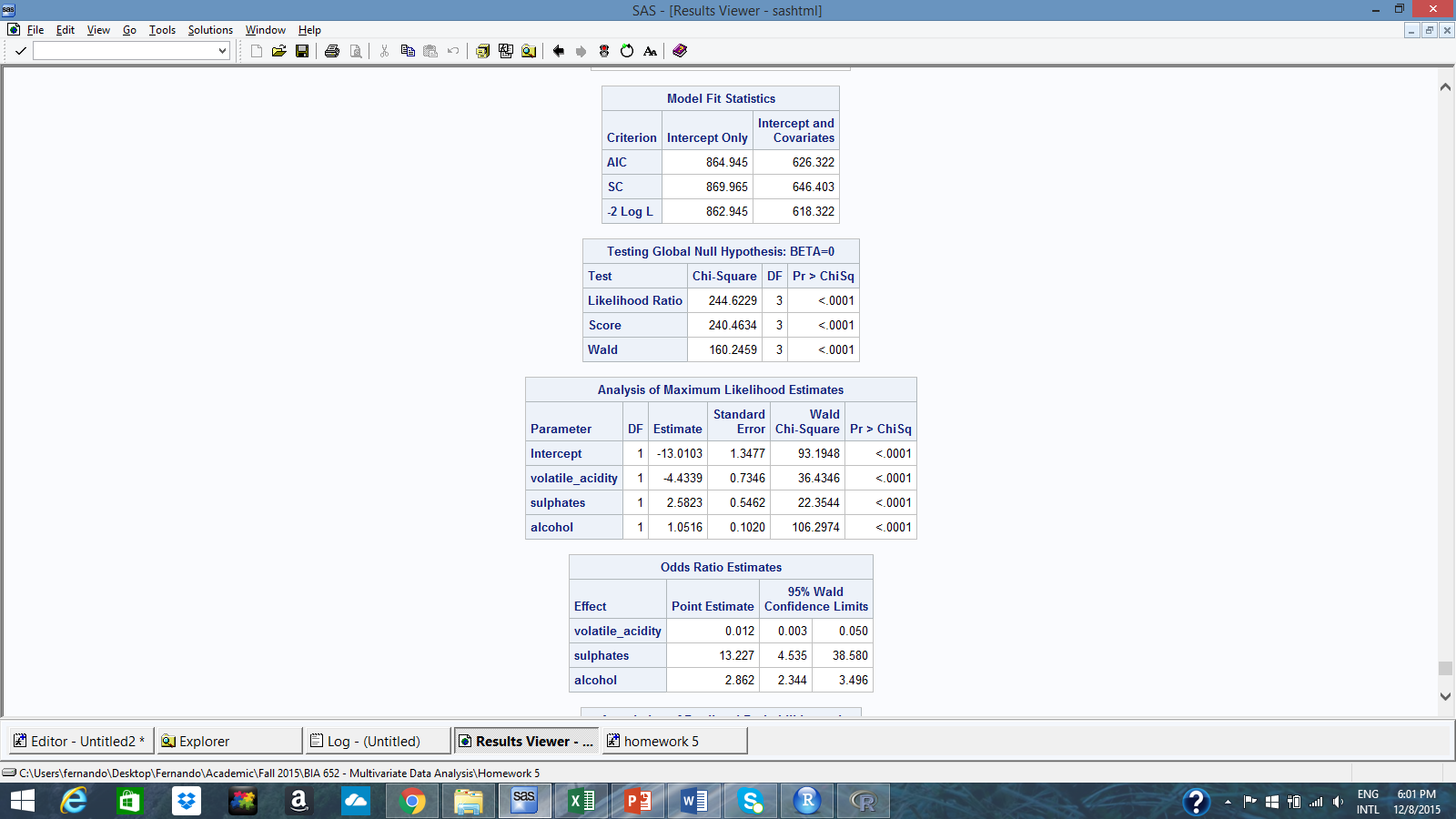
* Alcohol
* Volatile\_acidity
* Sulphates



**Modeling – White Wine**

In order to model the white wine, we performed logistic regression on the white wine dataset. The following variables gave the best model:

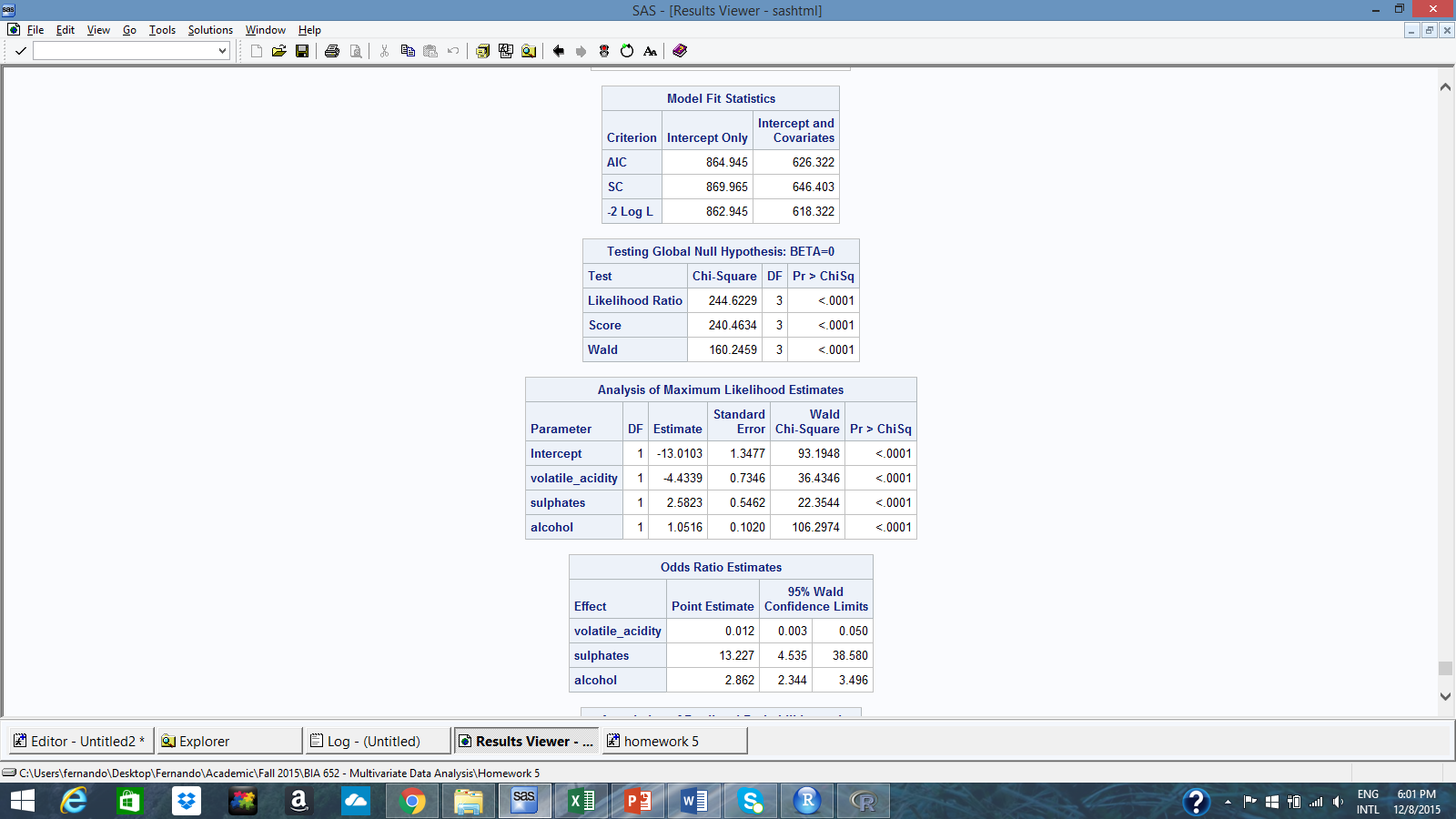
* Alcohol
* Volatile\_acidity
* Residual\_sugar
* Free\_sulfur\_dioxide
* Sulphates
* pH



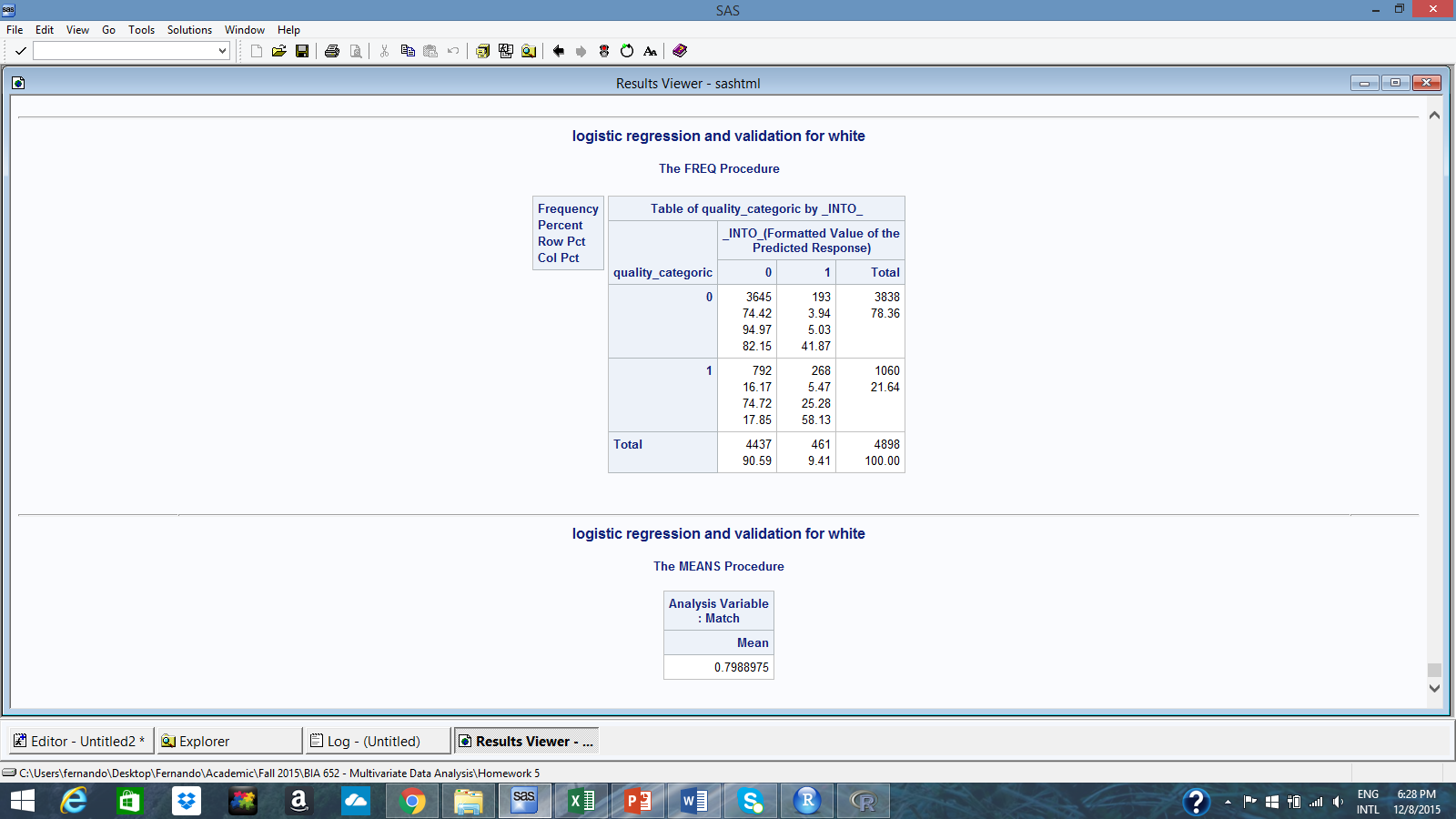
**Modeling – All dataset**

In order to model the All dataset, we performed logistic regression on the All dataset. The following variables gave the best model:

* Alcohol
* Volatile\_acidity
* Residual\_sugar
* Sulphates

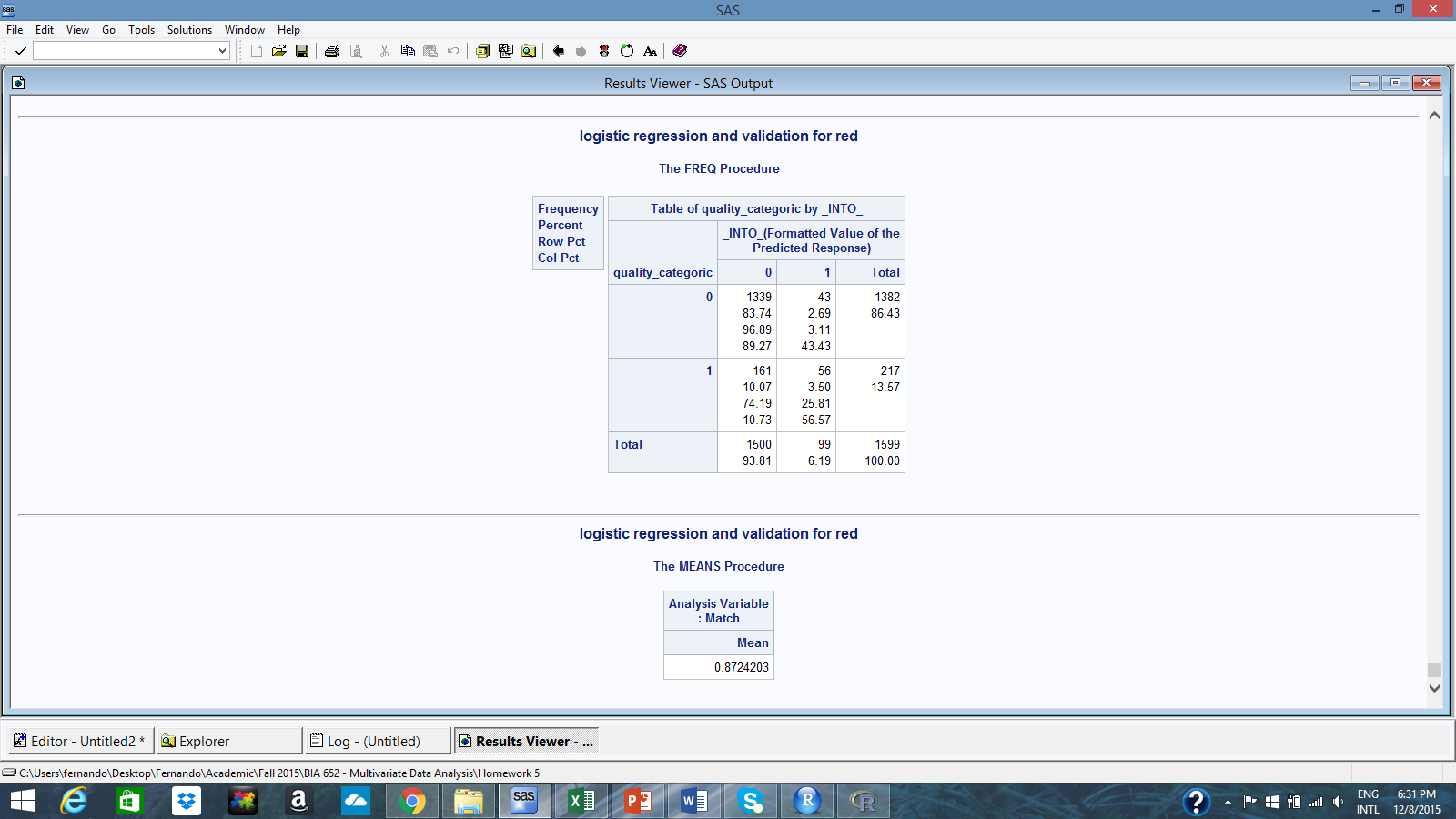


**Evaluation – White**



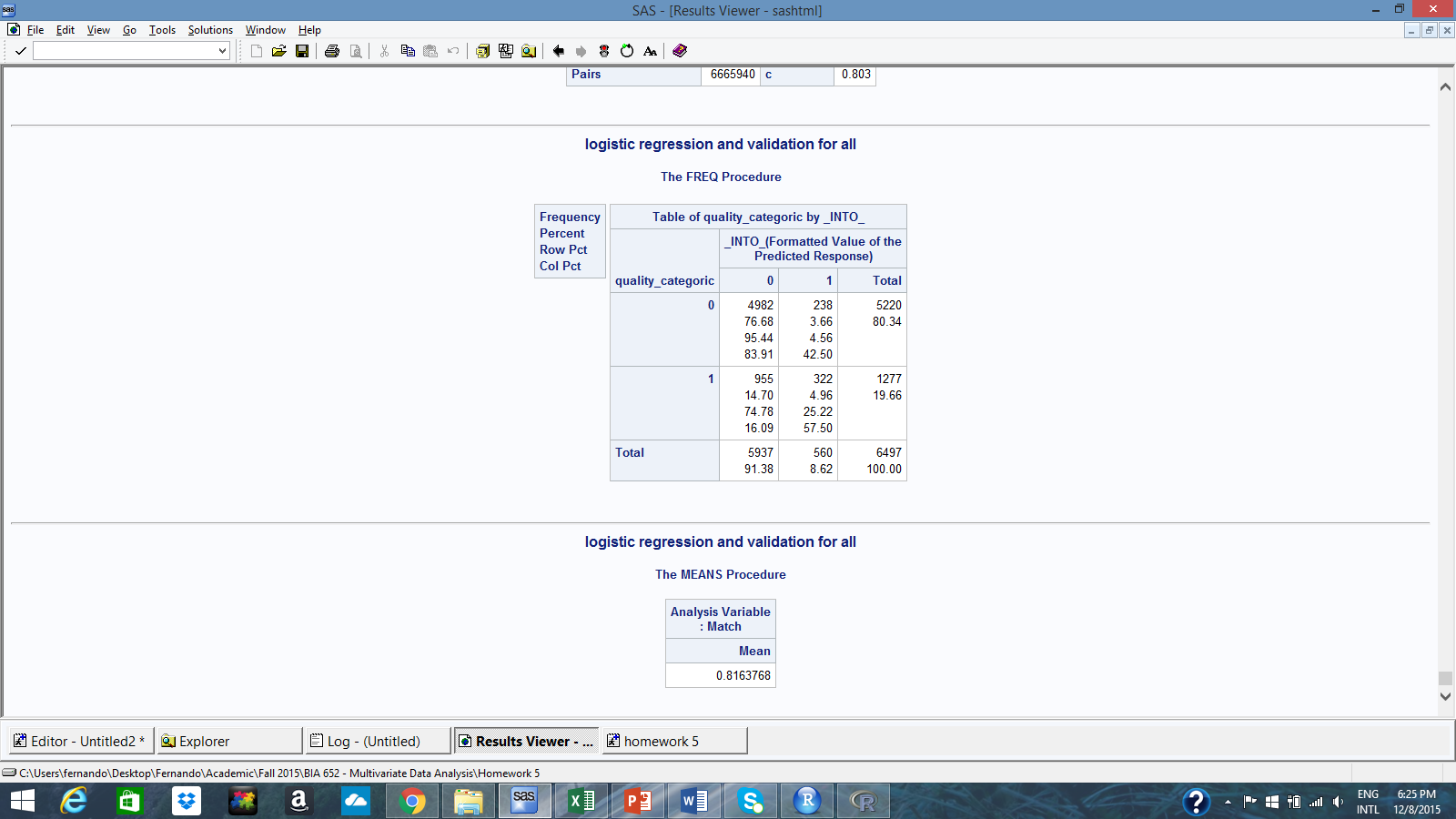
Analysis: 79.89% of good predictions!

**Evaluation – Red**



Analysis: 87.24% of good predictions!

**Evaluation – All**



Analysis: 81.63% of good predictions!