UNIVERSIDAD POLITECNICA DE YUCATAN

QUARTER No. 3

DATA PREPROCESSING

UNIT I EDA

RED WINE QUALITY EDA

STUDENT: HERNANDEZ MARTINEZ ANDRES ALEJANDRO

16/10/19

GROUP: DATA 3B

Content

Introduction

In this project we will be analyzing the data set, with the help of some libraries, statistical knowledge, and programming skills. The goal is to present a descent EDA with the essential, this with educational purposes, we will be exposing code and plots to help the comprehension of the topic, and of course, of the EDA.

An EDA with good quality is going to be helpful for further investigations, and in a shorter scale, for the understanding of the career. In fact there are many factors that would alter our EDA and as long as there are no templates for an EDA, nor an specific guide to make an EDA for every Data-set in the world, but there are some guys that dedicate a part of his time to make this EDA's so all the people can based their future projects with different analysis and way to interpenetrates.

There will be some problems like the duplicated values, where we will be taking decisions, and we will need to considerate the consequences of those decisions and make a balance if it is convenient to keep them or drop them. Or the standardization of our categorization variable that we will be meeting later. We will try to get some significant tables, where the information is well organized, and that they mean something. Also with the help of some methods we are going to give important information about the essentials of the

Starting with the analysis.

To start with this analysis, we will have to take in mind what we are doing, besides, somehow we need to represent our data in a readable way, and with useful variables, an information to use it later in the EDA or in a deeper analysis. There are some trivial steps such as uploading of the file, that some people consider not important to mention, but as long as I want this document to be a complete, I will try to specify every step of my EDA.

- Getting the .csv

We can get the csv from:

'https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv'

In the UCI page, we can get a great variety of Data Sets that can be very helpful, when you are a beginner, cause' these are very well formatted data, with almost no problems of structure, besides they are typically complete. So now that we know where we got our dataset, we can continue with the first part of the EDA.

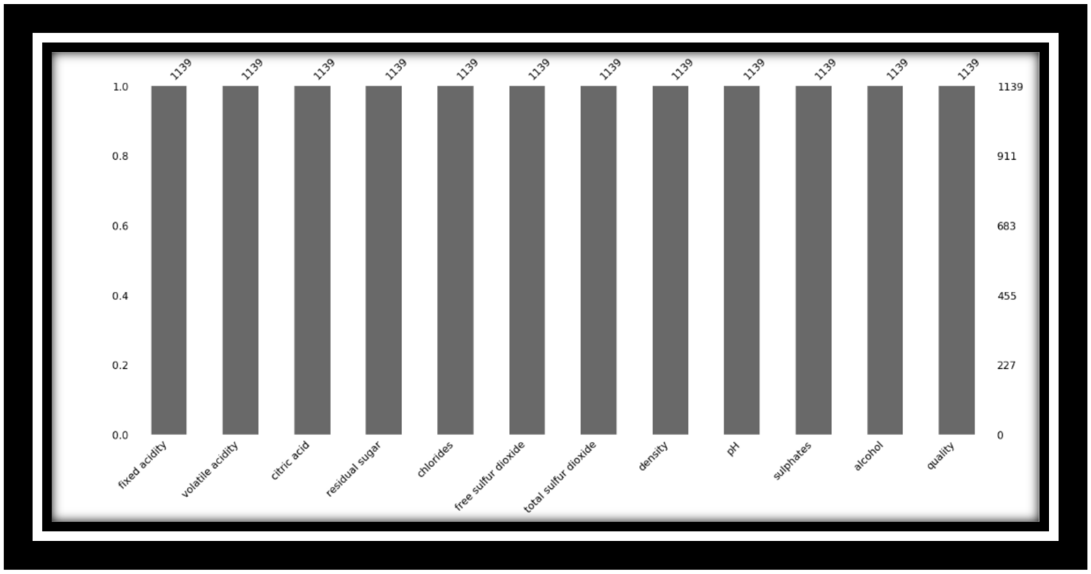
- Meeting the Dataset.

In this section we are going to explain the data that we will describing the Data set with the help of some tools given by the 'pandas' library and some plots in order to have a visualization of the dataset more graphically.

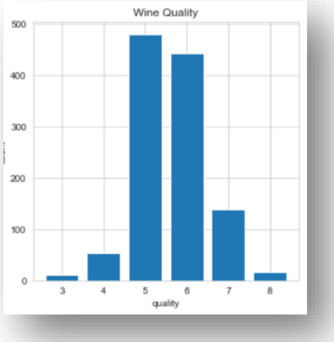
- **Description:** We will present the raw Data set, information such as the name of the columns will be stored into variables to later uses, they are going to be important if we want to plot, or we want an specific columns. The .shape function will help us in order to se the registers that we have in our data set, basically to dimensioned the Data set.
- **NaN values:** The NaN values would be a problem in the case of their existence so, it is important to check every time you get any data set if you have your values complete, otherwise you might want to chop, or change their value.
- **Duplicated Data:** Another problem that we might want to eliminate in the case of the existence, but there are some cases where your criteria is going to be the one that decide if they stay In the Data frame, or no, cases such as general description where the variables can be repeated, is important to keep them, otherwise (like our case), it is important to eliminate the cause they do not contribute nothing more to the analysis.
- **Categorical Unique Values:** It is important to take in mind what are you analysing so, to check the values that a categorical variable can take could be very convenient, besides you will be seeing these variables in your

Categorical vs Continuous plots, and they will be target of some of your matrix.

- **Distribution of the Categorical Values:** To see the behavior of the dataset, such as the count of an specific variable and presenting it in the plot would be convenient If you want to see if your Data frame follows any distribution, or it will be necessary to apply a transformation.
- **Display statistical information:** In this section is where we start to make interpretations, and in some point an analysis, but as long as the '.describe()' only gives us a part of this statistical information, we need take use of specific methods to get the information that is missing such as skewedness, kurtosis, mode, and median. Once we have this data in a working with the plots will be much easier, because, it will be the representation of these numbers



Observations: In the plot above I am basically checking if my dataset is complete, in case we are missing some values in any of the variables, this graph should tell me in which of the variables is this null value, and how many are there



In the graph on the left, we can check the count of the qualities, so basically it has a normal distribution, as you can see is very contracted, and there are not values that are skewed

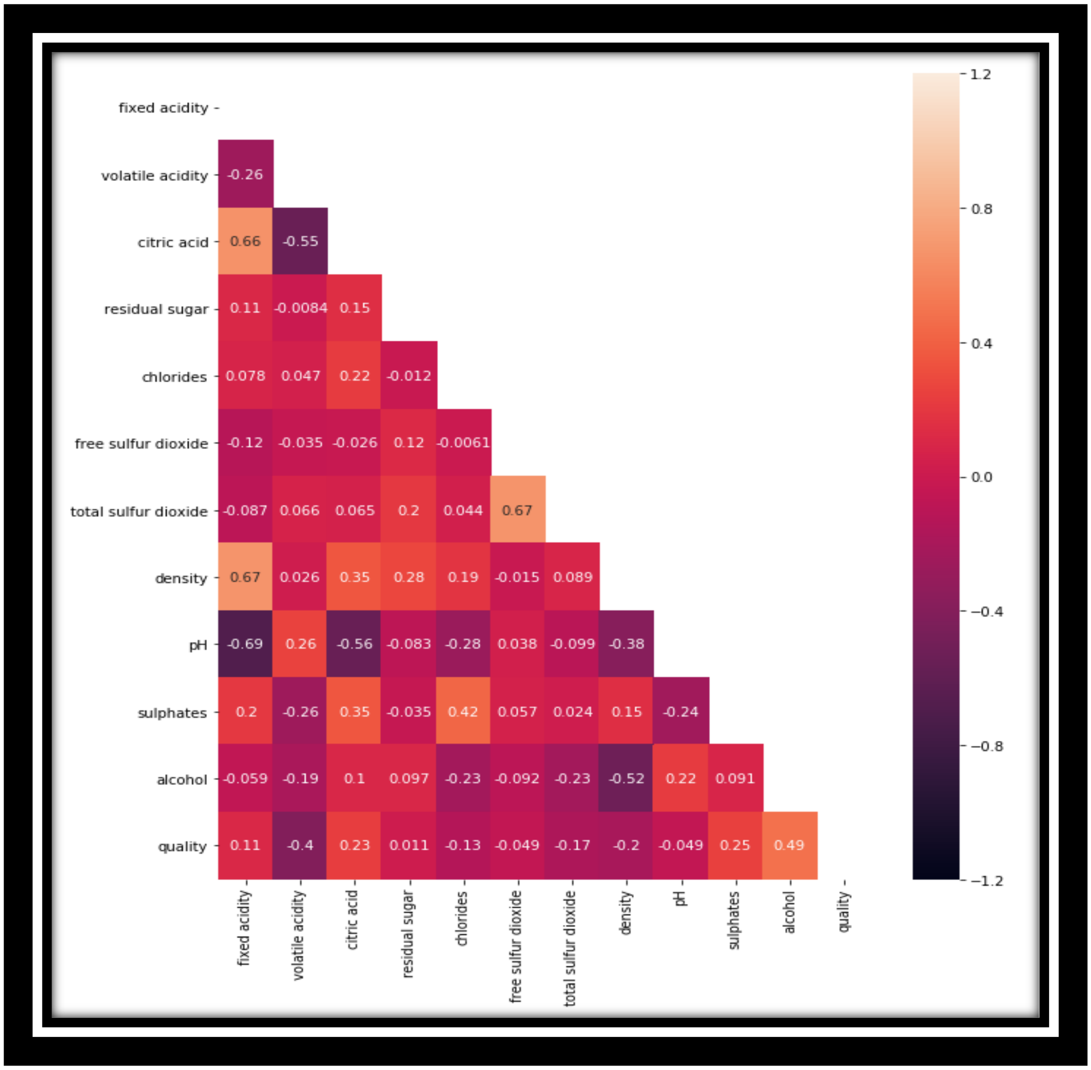
Plotting analysis

In this section we will be plotting important graphics to understand our data frame, the variables taken for the comparison will have sense to the previous information. As we know there are 3 types of analysis when we talk about plotting: Multivariate, Bivariate, and Univariate. So, we will try to cover these three points with plots that have enough sense for the next person in production. A brief definition is going to be mention in analysis that we give.

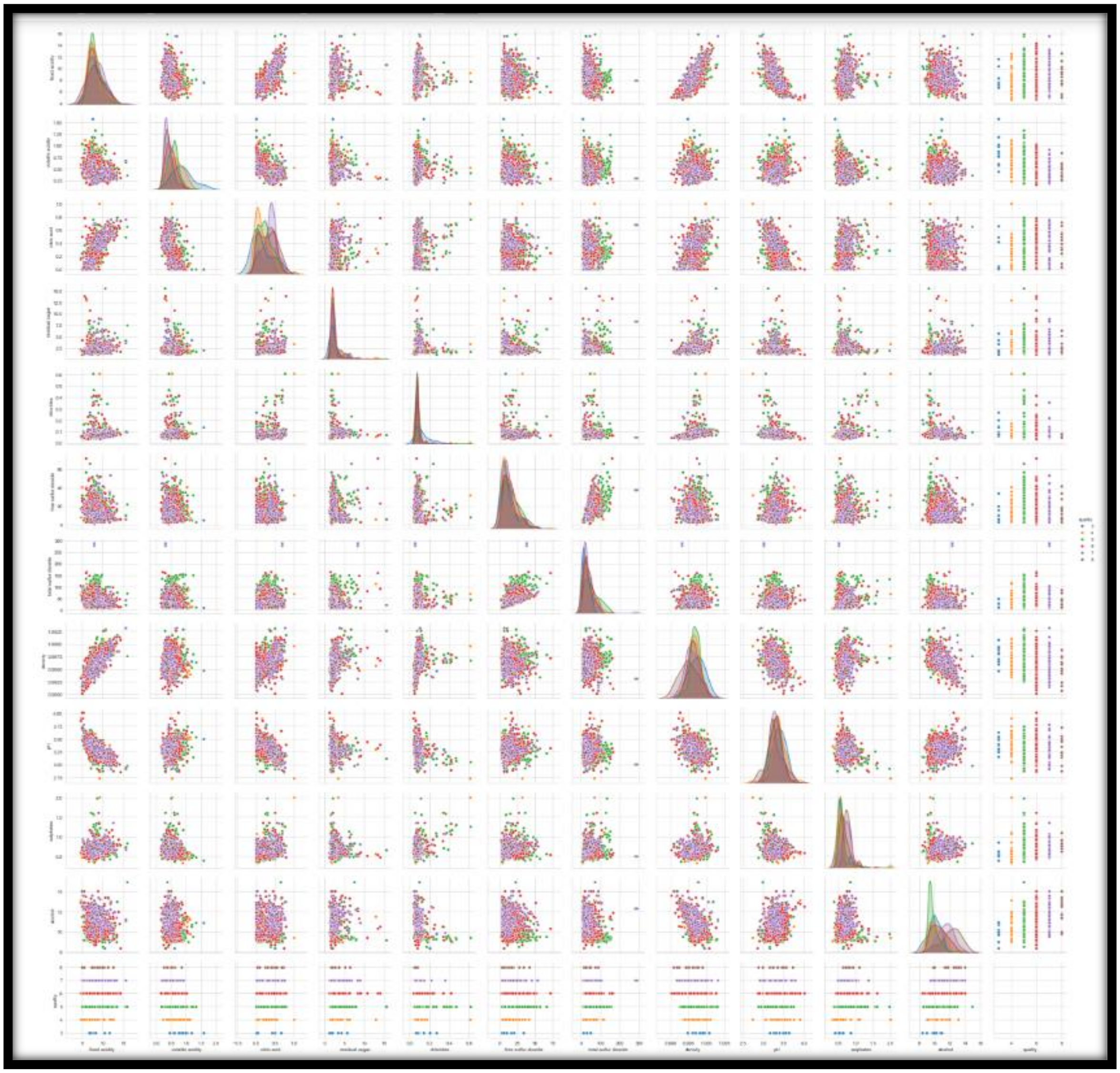
- Multivariate Analysis.

For the multivariate plots I made use of the correlation matrix, and the scatter matrix, and the covariance table. These because they made a lot of sense with the data that we had, and these graphical and no graphical resources are helpful to clear up our analysis, and to see the variables that we are working with and the behavior of each in a graphic and no graphical way. The consideration of 'Multivariate' analysis cannot be missing from an EDA of this dimensions, in the case of less data, you might want to have the correlation matrix, just for an overview.

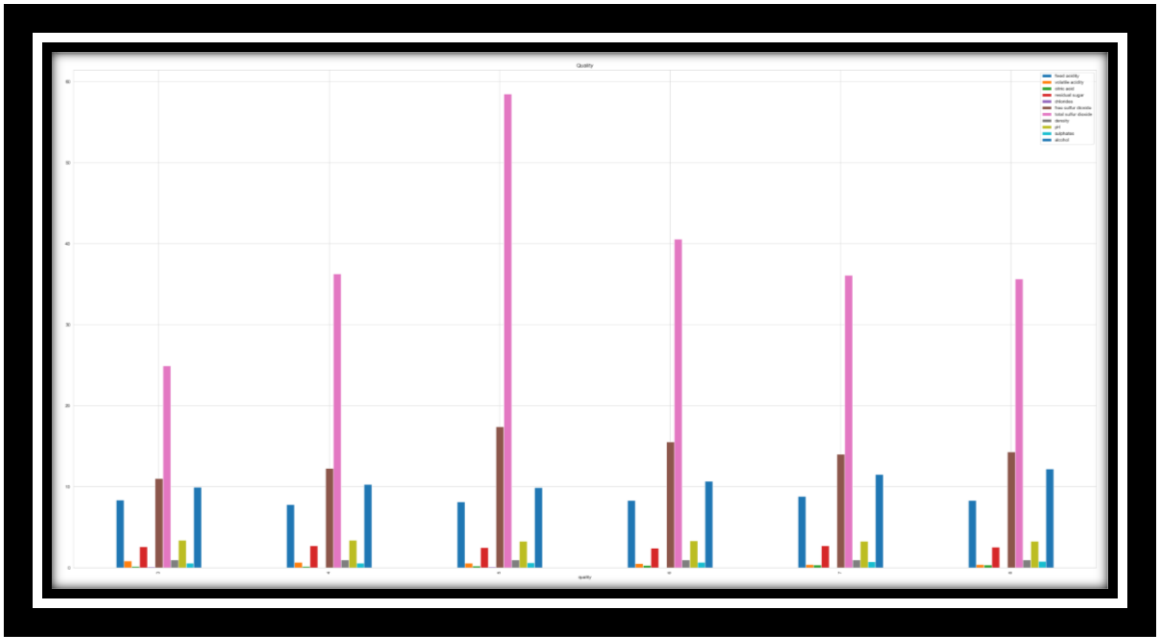
- **Correlation Matrix:** In this matrix we want to see which variable are the ones that are more correlated, as you can see in the graph we have in the right an scale of colors that will tell us the color according to the correlation coefficient, the most they approach to the limits the highest is the correlation, so it is not surprise that all variables are perfectly correlated with themselves, that is why we up to put a mask in that part of our table, it is redundant now that you know that there is a perfect correlation with itself. As you see in the matrix, there are 4 pairs of variables that are “Highly” correlated, which are the ones that are with a darkest color, and the correlation between quality and alcohol. So, from this table we have an approach to the variables that we my be interested in. I can conclude from this matrix that there in general the correlation is quite week but with the classification that we mention before we have something to work with



- **Pair Plot:** In the pair plot we will be analyzing over every single variable of the data frame putting as reference our target quality which is going to be plotted in all the graphics that are inside the pair plot, this graphic is going to be very helpful because from that plot we can also choose the variable that we are going to be working with, besides, it gives you a visually supply of the correlation matrix. From this graph I can conclude that we can make some regression process to a couple of pairs because of its behavior in the plot, besides that I can get a plot that is going to help me to classify that is the relation of the alcohol with the quality.



- **Bar Plor:** In this plot what I did was to group my data (in a new data frame) by the quality, so as long as it is my only categorical variable, I would say that has sense, besides, if you want to compere the values of each variable, you have solved this problem with the trable, but mor related to the EDA, I chose to make this table to dennotate that it is hard to clasificate the data frame, and to get a prediction would be very complicated. So as I said is very difficult and risky to conclude something from this graph, so the observation here is that the data is very likely in reference of my categorical variable, and it is going to be hard to categoraize the next wine in the list into any of the califfications.



- Bivariate Analysis.

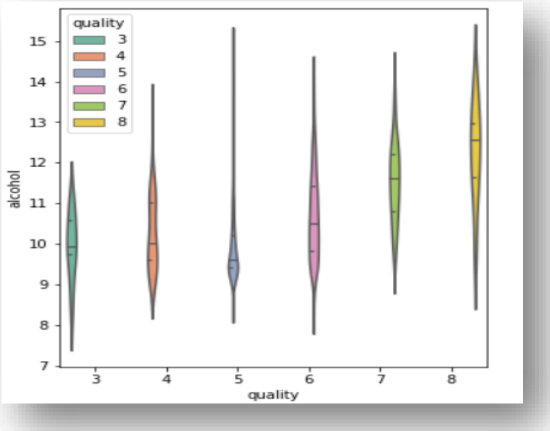
In the bivariate is when you basically have (as the name say) two variables, and only two, so basically is to make a close up to an specific behavior that we may want to represent into a plot, also, there are some tools that we are going to be applying, such as the linear regression that will have sense in our EDA because it can give us an approach result to next registers. Bivariate analysis can be divided into different classifications, but the ones that we are going to be applying are going to be:

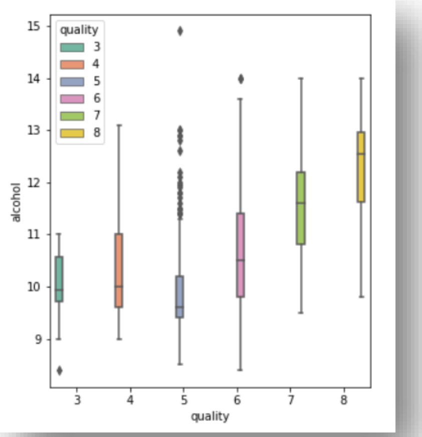
- Categorical x Continuous
- Continuous x Continuous

✓ *Categorical x Continuous*

In the Categorical x Continuous we will have a categorical variable as target, and a continuous variable, this type of plots are going to be very helpful to see the behavior of an specific variable, and in case of classification it is very helpful because you may be able to see the different distributions accumulated in different parts of the graph.

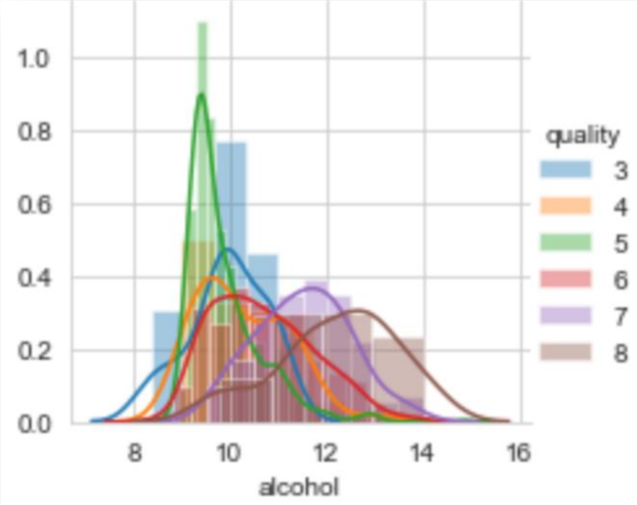
- **Violin Plot:** In the violin plot we will be able to see the distribution that the variable has in relation with our target, it will always be necessary to have this target because is going to give us the insight of the variable. It gives you just an idea of the outliers but not very specific. From the graph next to the text I can conclude that the data is dispesed, and that there isn't a notable aggrupation data in any of the quaritles, but I can say that the lowest the alcohol the lowest the calificaitons in most of the cases. And if you have more alcohol the calificaiton tends to be greater





- **Box Plot:** In the box plot the situation is that it will give you the insight of the continuous in relation with the categorical, but the result will be that this is going to give us the outliers marked and a well defined quartiles. From the box plot I can conclude that in the mean the of the quality are many outliers, and that the classification of those wines may not be correct, or there are other factors that alter the decision. But as long as it is the one that has more differentiation between the variables, they are just outliers.

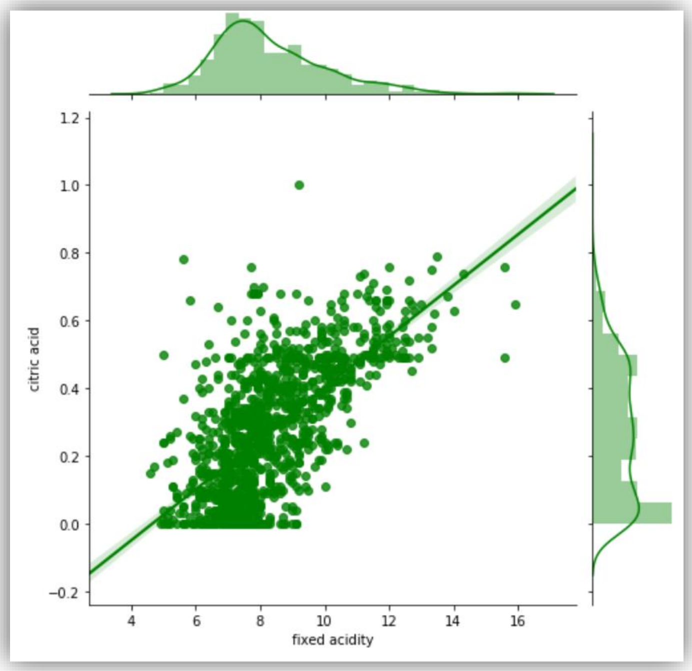
- **Distribution Plot:** Here in the distribution plot of the alcohol and the quality, seaborn gives us witch of the variables are in the distribution, so it is easier to find the differentiation. Then we can check also some of the outliers of the comparison, so basically this is the reason of the other two plots, because the other ones where going to give me a plot with almost no differentiation and the dimensions of both, the boxplot and the violin plot, where going to be almost the same



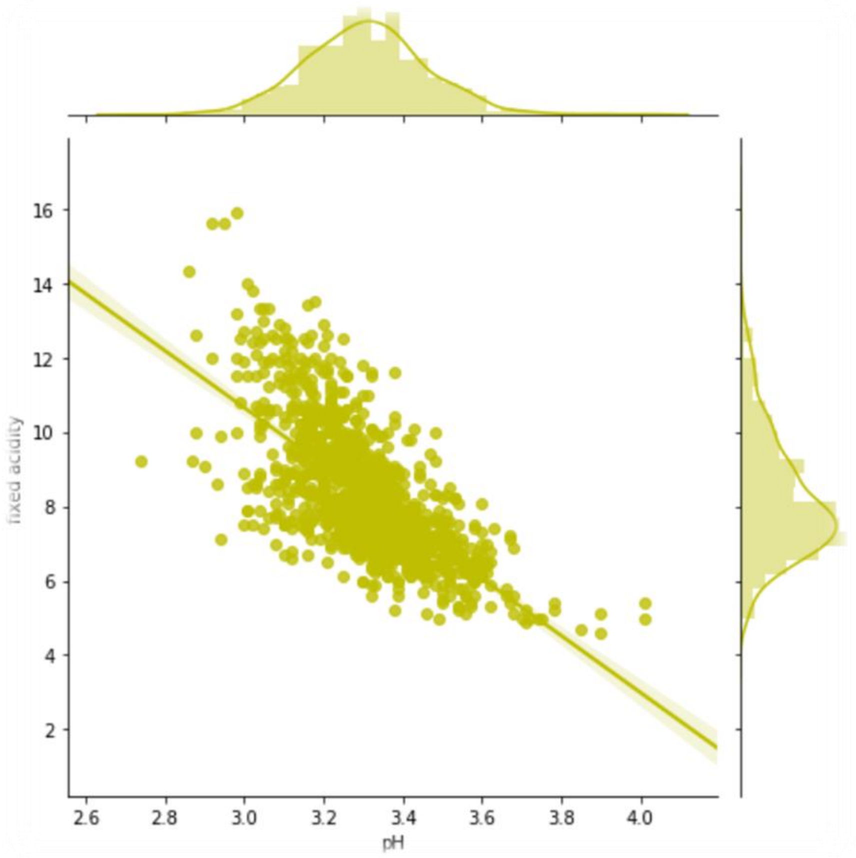
✓ ***Continuous x Continuous***

In the Continuous x Continuous graphs we will be matching two different variables that have no categorical properties, so here basically are the variables with numeric values, and the ones that can apply to make some math with them, that is why we need to be careful with the comparisons, we cannot make just comparisons because we wanted to make comparisons, here we will need some research to give sense to the graphs that we are going to show.

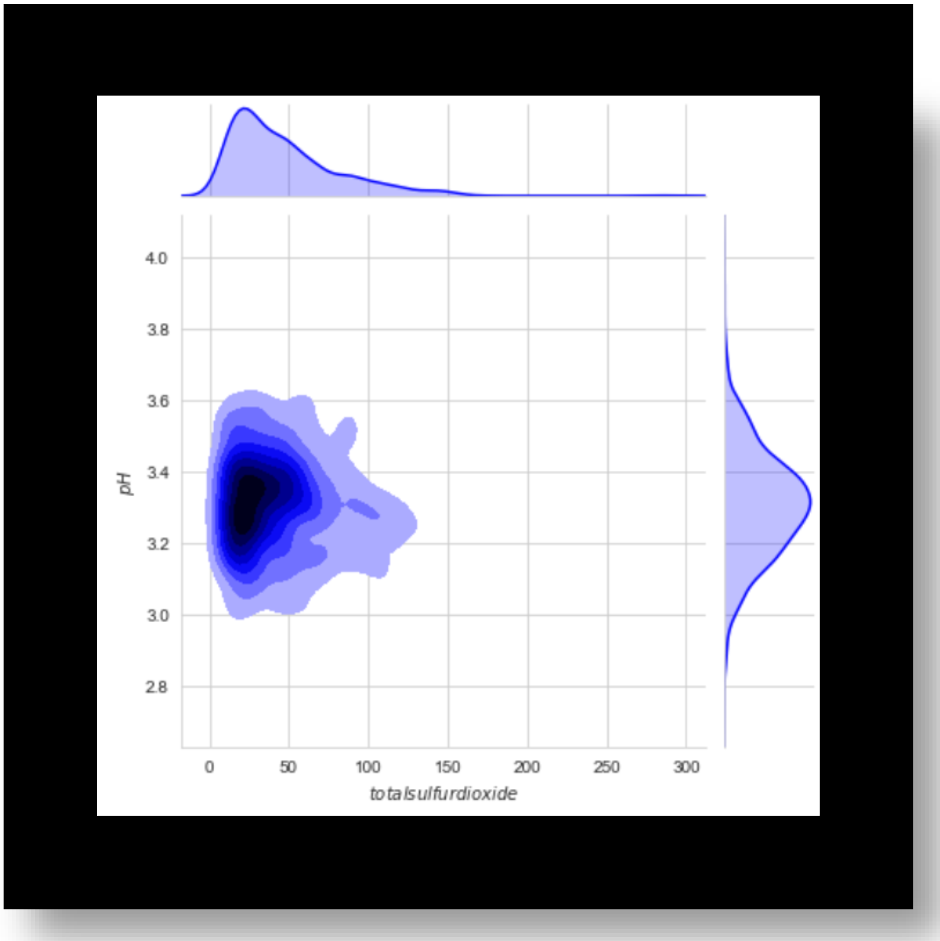
- **Linear regression with marginal distributions (1) :** This plot is going to give us a lot of information that is very useful in the analysis, it gives us the distribution, of each variable, the comparison, and the linear regression. The plot is original, but the idea of compering these two variables was inspirited in a Medium EDA, so basically, here the graph is telling us that the fixed acid is going to be growing as long as the citric acid grows, so basically the correlation of these two variables gets clear in this point because we can see how is growing concurrently. And the skewedness of the fixed acid makes it to be in the beginning all the values.



- **Linear regression with marginal distributions (2) :** This other plot is basically the same of the one above, but with two different variables, the contribution of the idea is from the same EDA, but well the explanation of this plot is that as you can see we have a, ‘normal’ distribution in the pH and a right skewed distribution in the fixed acid, this is why the accumulation of the points is in the bottom of the graph, and so the regression that has a negative slope, so as you can see, there is a depression in the values when the pH stats to grow, this is because a pH that is greater than 4 the pH starts to be basic, so there it is the opposite of the acid. That is why the slope is in depression



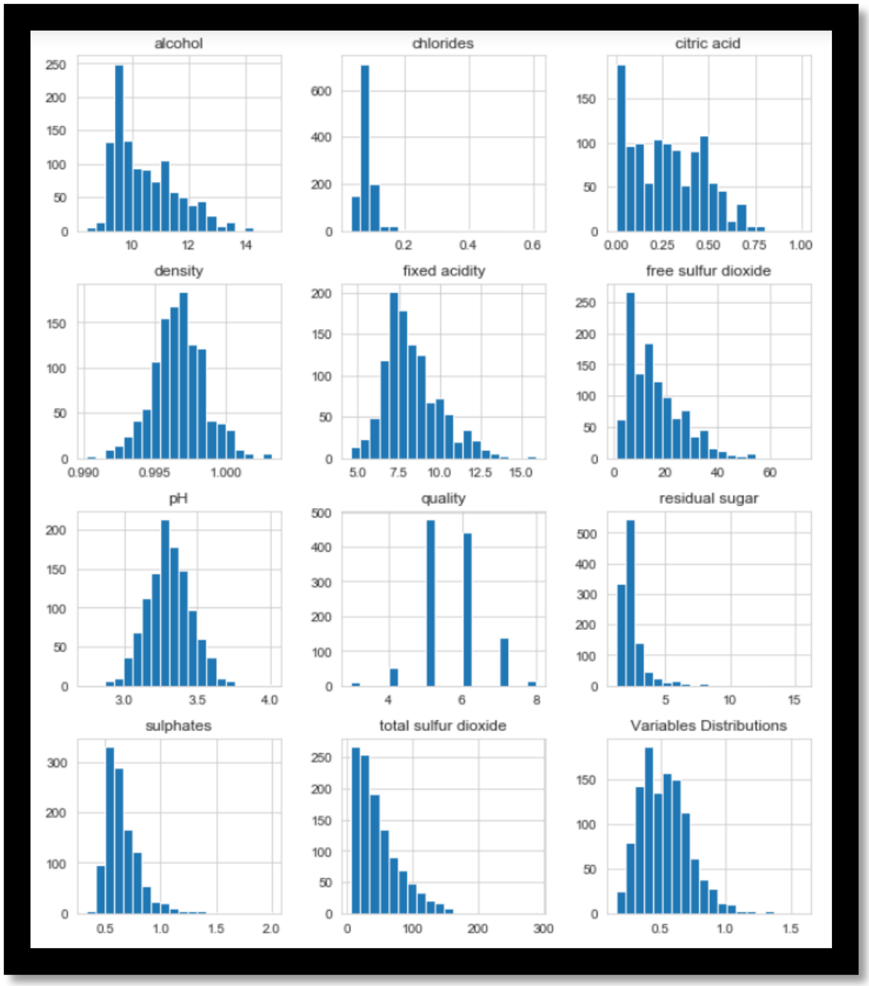
- **Density Distribution:** This graph is 100% original, this is because I started to analyze the variables and I realized that as long as the variables are chemicals they have pH, so I decided to choose (from the pair plot) a graph with a grate accumulation In order to verify that there is not change of the values, because of the standards of the wine. And what I found is that this data is from Europe because the standard is 50 ml and as you can see in the graph the accumulation is oscillating in the 50’s so we can ban that is an American data frame, because the standard in EUA is up to 350 ml.



- Univariate Analysis (Visualization).

In the univariate analysis the plotting is going to be much easier, and the classification, this because there are not lot of things to do in this data set (as far as I know), this because the other analysis gave us important insights. The univariate analysis its going to take only one variable in count so we cannot get very creative in this section.

- **Histogram distribution:** In the histogram distribution of each variable we are going to be able of the distribution that it has, the kurtosis, and skewness are going to be very present In this histograms, which is what we are going to analyse. We can see that al least a great part of the histograms are skewed to the left, but there are others like the pH that are close to a normal distribution, so the suphates variable is very close to the normal distribution, the citric acid distribution is harder to see but I would say that is a platykurtic, and in the oposite side, the chlorides is a leptokurtik distribution. The other ones is claear that are skewed



Conclusion

From the result that we got from this data frame, we can conclude that is not useful to make a classification by the quality, because a prediction would be very risky and would not be trustable. But if we don’t want to make a prediction but a chemical analysis over the data frame as the one that we made in the end, we might get some important behaviors in the chemical aspect of the wines, and characteristics that are constant and other ones that have some functions to follow, my lack of investigation, limited me to do fit a regression, but I feel confident with the insight that I got from the linear regression. But even when the only categorical variable was a little bit messy to work with, I got some information that can be useful to furthers analysis. I would say that the order is very important In EDA’s because it was easier to do the analysis when you have: the description of the data, with useful variables related to de data frame, and secondly the multivariate plots which were my guide in order to choose the variables, that were more easier, and worth to analyze. I focused on the variables that my own EDA was telling me to focus, because of its distribution, and behavior of the comparison.

We need to analyze the data frame deeply in order to find information such as the hypothesis that I propose, where I said that the data set was drafted form Europe because of the properties where they have restrictions, and I’m sure that there is a lot of hidden data that we could get with a deeper investigation. If I was sure of my hypothesis, I could have make an analysis of which of the wines are attached to the normativity that Europe has, but for matters of time is not going to be possible to be explained in this project report.

References

Dave, A. (21 de 02 de 2019). *Medium*. Obtenido de Regression from scratch — Wine quality prediction: https://medium.com/datadriveninvestor/regression-from-scratch-wine-quality-prediction-d61195cb91c8

Gottipati, S. (30 de 01 de 2019). *Medium*. Obtenido de EXPLORATORY DATA ANALYSIS(EDA): https://medium.com/@srivathsagottipati/exploratory-data-analysis-eda-4b81d84ef5cf

Patil, P. (23 de 03 de 2018). *Towards Data Science*. Obtenido de What is Exploratory Data Analysis?: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15