Assignment A2-LP3: Data Exploration and Preparation			
Student Name	**Shantanu Gupta**	**Student No**	*218200234*
My other group members		**A2 Group No**	*As per CloudDeakin group number*
Team Names	**(as per record)**	**Student Nos**	*Student number*
	(as per record)		*Student number*
	(as per record)		*Student number*

	Exceptional	Meets expectations	Issues noted	Improve	Unacceptable
Problem Statement					
Explore & Prepare Data					

Brief Comments	**Read these notes as we are really trying to help you out!** **Remember: If it is not in this report, it does not exist and does not get marked!** You can use the above form to estimate the expected mark against the rubric (see the assignment "info" document). Be realistic and note that we will find many problems you may not be aware of. Assume that markers may be tired when assessing your work and they may miss some important aspects of your submission when not presented clearly, or when you deviate from the structure of this template, or if you do not include them in your report. So be clear, number all tables, charts and screen shots used as evidence, describe all visuals, cross-reference your analysis with evidence. Submit this report in PDF format to avoid accidental reformatting of the content. Submit all RapidMiner processes (.RMP files) in a separate ZIP archive, so that if there is any doubt we could load your work and replicate your results (we will not do this to find missing report parts). Ensure that the report is readable and the font is no smaller than Arial 10 points. In the report include only the most significant results for your analysis and recommendations. You will be able to submit your work once only so make sure you get it right – check these before posting on CloudDeakin: Is this your document? Is this the correct unit, assignment, year and trimester? Is your name entered above? Is the group number included and is it correct? Are names of your group members entered as well? Are all pages included? Does it all fit into the required page limit? Have you zipped all RapidMiner files (.RMP files)? Is the report contents yours alone? Then after the submission – check these: Has the PDF report been submitted? Has the Zip archive of RMP files been submitted? Can you retrieve and reopen both back from your submission folder? Note that the late penalty will be calculated on the date and time of the last submitted file. Finally, as all reports will be inspected for plagiarism, ensure that your analysis, your evidence, your way of thinking, your report and its presentation are unique and demonstrate your ability to create it all independently. So if you work in a team compare your submission to those of your team members and make it quite distinct in both contents and form. Any part of this report that bears any resemblance to another students' report or any information source written by others or by you for another unit (e.g. on the web) will be treated as plagiarism.
Total	

Executive summary (one page)

Expectation

Australian Wine Importers (AWI) give data that contains 130k of wine reviews, where each description review is written by a sommelier, and give an 80–100 point rating. Wine Enthusiast ranks wines on a 100 point scale with only 80+ point wines receiving a written review. According to this blog post, the scores roughly correspond to:

- Classic 98–100: The pinnacle of quality.
- Superb 94–97: A great achievement.
- Excellent 90–93: Highly recommended.
- Very Good 87–89: Often good value; well recommended.
- Good 83–86: Suitable for everyday consumption; often good value.
- Acceptable 80–82: Can be employed in casual, less-critical circumstances

Since each wine should have been expertly described by someone trained in the art of wine tasting, AWI thought we could use the data to analyse wine description. In Wine Reviews, the wine description plays a vital role. The "description" of each of these reviews is comprised of a few sentences in which the wine reviewer describes what he/she tastes in the wine and gives an opinion about the wine. A good description can make your wine stand out. It also helps get reviews faster. The objective of this project is to build a machine learning so that we can build a predictive model to estimate (rating) of a wine belonging based on descriptive words? Lastly, it will help you get some points and sound smarter when I review my next one. Let's see what we can find in the wine description. We will use Rapid Miner for handling this project.

Extension

The dataset was already very clean, so there wasn't much I needed to do to it in terms of pre-processing. There were no strange symbols or HTML stuff we needed to strip away. However, there are a couple of things I did need to do before giving to clustering.

- I ran the descriptions through TF-IDF Algorithm to find words which were useful in classifying the different varieties. The TF-IDF score finds frequently occurring words but down weights them if they occur often. For example, if "fruity" occurs in 99% of wine descriptions, then it is not a very informative word and consequently down-weighted. Other words like "tobacco" may only appear in a subset of fuller body wines and would be useful in classification so it receives a higher weight.
- Transform cases-To convert all text into lower cases. It's no matter whether the word is in lowercase or uppercase.
- Tokenizing cuts down sentences into individual words. Involves breaking down the text into individual words and trying to find patterns in those words. For example, are certain words more associated with a particular varietal, or maybe with just red or white wine?
- Stemming cuts down words to their root. Reviews may contain words like "fruit," "fruity," or "fruitiness,"—and those words would be counted as three separate words when really they all mean the same thing. Stemming cuts them down to the root, "fruit," and counts them as just one word. We used the snowball method for stemming.
- Then we filter stop words in English and remove those words that have a length less than 4 and greater than 25. We can see a huge difference after normalizing our text. Now we can see our text is more manageable.
- Next step, is to select top 50 attributes on the basis of weights. After that, we passed through the clustering process.
- I decided to apply the K-Means clustering algorithm with default values, and it basically divided everything into five clusters. Already we can see some patterns in the description (or even the words) based on these relationships. The chart is great, but what do the cluster really mean? We will explore in the next part.
- After that we apply different machine learning algorithms like decision tree, neural nets etc. After that we optimize the parameters and create a final deployment process.

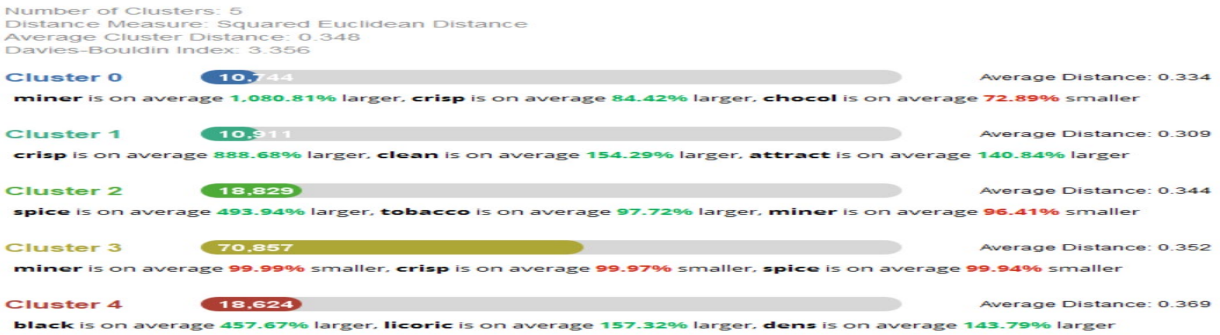
I have explained the data pre-processing steps in Extension section of Data exploration. I have explained all the steps through description column but this step is similar for structured data.
Unstructured Column-description, Title
Structured Column-Variety, winery, country, province

References:-

The official 2019 Wine Vintage Chart (2019), https://bit.ly/2F7CUDH
Exploring and Classifying Wine Enthusiast Reviews, John, 2017, https://bit.ly/2VBNfk3
Introduction to Natural Language Processing (NLP), Morgan, 2018, https://bit.ly/2UNfTtZ
Predicting Wine Quality using Text Reviews, Feb, https://bit.ly/2ZJrjTt

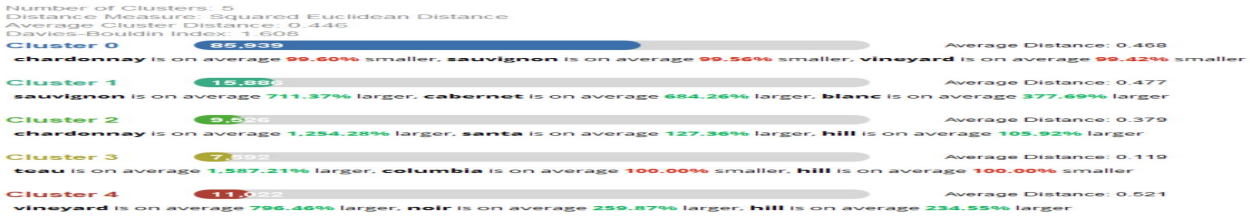
Data exploration and preparation in Rapid Miner (one page)

Expectation



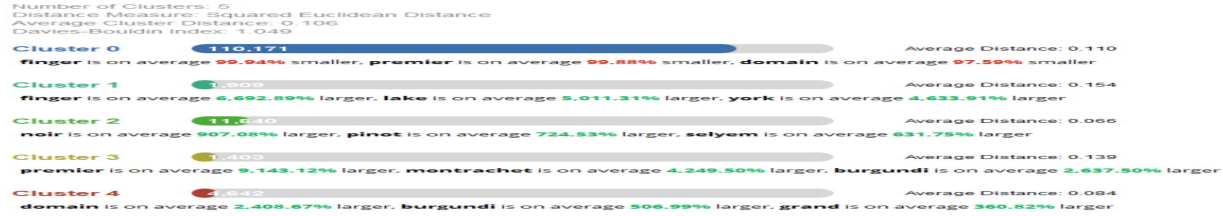
Picture 1: Clustering Analysis on Description

From picture-1 cluster 3 is the most important one as most of the words will fall into this category. As we see in cluster 0 the word miner and crisp have appeared the most number of times. But that same word crisp also appears in cluster 1. It is possible that the same word occurs in multiple clusters with different probabilities. The green colour representation shows a high possibility of the word occur in that cluster while red representation shows that word have very less chance to appear in that cluster. So cluster 4 contains the words like the black, bitter, or sweet taste, cluster 0 contains words like power, texture, melon or chocolate. As you see that cluster 0 contains taste words, cluster 2 contains spices words while cluster 0 and 1 contain more words towards texture and concentration. In every cluster, it is showing the top three most common/frequent words or less/rare words. In conclusion, we can see that what is the most common words or rare words used by the reviewer in the description of wine in social media.



Picture 2: Clustering Analysis on Title

I have done a similar analysis of the title column. The analysis of this graph is telling us which word is most significant to the clusters. Red means the words used rarely while green word means more frequent words.



Picture 3: Clustering Analysis on Structured Column

When we consider clustering analysis on the structured column, I used country, province, title, variety and winery as a column. I haven't used region, designation and taster name as these columns contains lots of missing values. So I have dropped them as it is not useful for the final model. As we can see cluster 0 is the most numbers of the words are there. As we see in cluster3 pinot, premier and province are the most important words. According to an analysis of pic-3, in terms of structured columns (description or title), I think new wine is most similar to cluster0. It will contain some of the words, reviews from cluster 0 as well taste also similar to the wine that falls in this category.

Extension

Missing values in numerical-Impute missing values by using Gradient Boosted Tree model. As our dataset contains more categorical columns, I used to replace the missing operator in which missing nominal values changed to average values (mode) of that column. I tried to impute the missing values using k-nn model but it throws memory error. As I delve deeper I found that country, province, title, variety, winery have very less missing values (In the range of 50-100). So there are two ways to handle it either by removing that row but we lost some important information, so it is better if we replace that missing value using average values (mode) of that column. Designation and regions contain so many missing values it is better to drop those columns. We reduce data dimensionality by using attribute-weight relationship graph. Either we can remove those attribute that is below threshold level or chose top 15 or 20 highest weighted words. According to an analysis of pic-1, in terms of unstructured columns (description or title), I think new wine is most similar to cluster3. It will contain some of the words, reviews from cluster 3 as well taste also similar to the wine that falls in this category.