Assignment A1-LP1: Data Exploration and Preparation			
Student Name	**Shantanu Gupta**	**Student No**	*218200234*
My other group members		**A1 Group No**	
Team Names		**Student Nos**	

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

Brief Comments

Total

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.

Executive summary (one page)

Expectation

Australian Wine Importers (AWI) asked us to develop a data mining method of classifying imported wines based on price category. Choosing a good wine for dinner based on the price has never been easier task. If we import a high price wine, and the Australian people doesn't like it then it degrades the organization reputation, a money-waster, but a meal spoiler as well. We need to come up with the top 3 or 4 parameters which can positively and negatively affect the price of our wine. We need to develop an algorithm to perfect the choice for choosing a high priced wine, without even tasting a single drop! And also ensure a quality final product, including organization reputation that lingers longer. This algorithmic method has better rather than relying on sommeliers, one's knowledge or an experienced wine person "recommendation" – all of which are expensive and sometimes unreliable.

Extension

If we further investigate the data, we can also identify the variety, winery, and location of a wine based on a description. We can also predict the number of points a wine will get.We could identify the wine based on a description that a sommelier mentioned or from wine reviews text.

Method

We accessed some 150,000 odd rows, on Kaggle wine reviews dataset. This involved analysing data in Rapid Miner. Machine learning methods were investigated in order to use predict prices, text mining for reviews and other attributes/factors to predict import better wine from the world in the Australia.

- Data Analysis
  1. Explore the dataset.
  2. Explore features in statistics way. Example-Mean ,median ,quartile ,box plots
  3. Then we will see duplicates and missing value
- Data Visualization
  1. In this stage our goal is not only to explore our data in order to get better predictions. We also want to get better understanding what is in data and explore data in 'normal' way.
  2. Common type of charts are:-Histogram, scatter plot etc.
- Feature Selection
  1. On this stage we want to make our dataset smaller without losing accuracy of model.
- Performance, Model Selection & Tuning
  1. Different algorithms are tested, evaluate the performance of the model and scoring of the model.
- Deployment/Production
  1. Final product in the market.

The bottom line is, we need to understand, manage and control the inputs and processes so as to ensure that we get not just the good quality of the wine, but it also should be price effective and reduce some kind of dependencies over sommeliers.

Bibliography

RapidMiner Studio: https://rapidminer.com/products/studio/
Kaggle (2018): Wine Reviews. https://www.kaggle.com/zynicide/wine-reviews/home

Data exploration and preparation in RapidMiner (one page)

Expectation

1. Except price and points all other attributes are nominal. Only Price and points are numerical.
2. We are using Price as label because our main problem is to import wine from other countries and categorize into 5 price range classes.
3. Other than price all our variables/attributes can be used as predictors. Degree of Dependence of the other attributes on the price variables is different. To test which attribute have most impact we need to use linear regression. Price may be depend on winery, taster name or may be country.
4. Correlation between points and price are positive correlated but it is not a strong relationship.

Attribut...	points	price
points	1	0.416
price	0.416	1

Table 1-Correlation Between points and price

5. The highest values isn't of the wine with highest points. The most expensive wine have points between 87 and 90

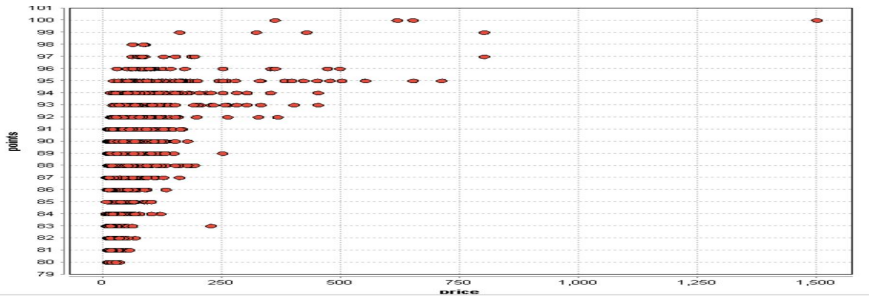

Figure 1-Graph between Points vs. Price

6. We can see that all wines have number of points above 80. And points has normal distribution. The most wines have 88 points.

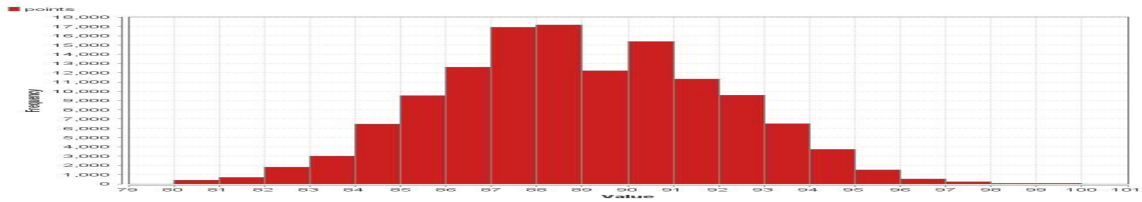

Figure 2-Histogram of points attribute

7. U.S.A. is main supplier of wine, they have the most wineries in the world followed by Italy and France.

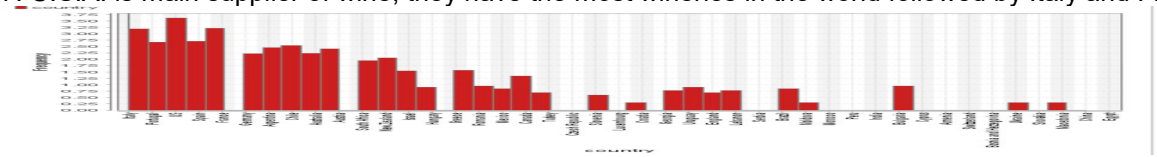

Figure 3-Histogram of most wineries in the world

8. What are most common words used in designation and description column to make a Word cloud.

Extension: Dealing with Missing Data

1. Predict with 100% accuracy-In the country column there are 63 missing values are there. We can predict these countries by using other column like province, region_1, region_2.As lot of missing values are in the province, region_1 and region_2, and we can predict these values from country column.
2. Leave record as is-We can do in taster-witter-handle column. As it may be possible that the person doesn't have a twitter account. We can also apply same technique on taste-name.
3. Remove record entirely-Title, winery, and designation column can be removed completely without affecting our result.
4. Replace with mean or median-Points and price missing value can be replaced with mean or median. For price it's better to do median as it doesn't affect by outlier. For points it is better to replace missing value with mean value.