Task 2

Business goals:

We are 3 students who are taking the course Introduction to Data Science. One of us is an IT major student and two are from chemistry major, which is why we chose the wine dataset - it is something that our team can understand better than a pure IT team. Since we are not doing this for any business, our goal is to just learn from the process and find interesting correlations between our data. Thanks to our chemistry background we can make some assumptions more easily and with less research than others, which opens up more possibilities for development. If someone were to be in the wine business and read our report, we hope to deliver them some sort of a formula and analysis of wine components and their relation to its quality. Our goals will be met if we manage to find correlations and create predicting models that other data scientists may not find.

Our situation:

For our work, we will be using jupyter notebook with python versions of 3+. The inbuilt and addable packages (like pandas, numpy, bioinfo kit.analysis) will also be used. We have at least 3 different PC-s in which we can work from and we are using an approach, where we all do our data researching, modifying and predicting first in our own-named notebooks. Then we analyze our results and choose the best one's for our final report. Since we have people from chemistry and IT, our mixture of ideas and approaches will probably give best results when combined and synced. We have already set up a github with mentioned notebooks and our main red wine dataset (a csv file). The dataset has also been verified. Our aim is to complete our work before deadlines.
There are not many problems which can cause issues for us, internet outages can be problematic, but even in that case, we can work in our own notebooks and collaborate when the data is back. Also, since all of us have 4G and a phone, we can still communicate with each other and present our ideas. Luckily, health related problems should be of no danger to us, since all of our work is done from homes anyway (so a quarantine does not limit our work).
The cost of this project for us is our free time and the benefit is being better in Data Science. However, if we find something interesting, this research might be something to lean on when looking for ideas to analyze later on. Also, if someone from any food or drink related industry wants to research some of their data, we already have some experience and ideas to help them with.

Terminology:

Algorithm - Repeatable sets of instructions which people or machines can use to process data.
Classification - the ability to use data to determine which of a number of predetermined groups an item belongs in. Will be used to predict wine quality.
Clustering - Clustering is about grouping objects together when there are no predetermined groups - objects are clustered together due to similarities between them and algorithms determine their value.
Correlation - Numerical relation between variables or in our instance columns. Shows us how likely a number from column A impacts a number from column B.
Data Mining - The process of analyzing a set of data to determine relationships between variables which might affect our outcome. This is something machines usually do on large scales, but we might be a bit wiser in chemistry to determine if something should be left out or not.
Data Set - the collection of data that will be used in our work. In our case, it consists of 1600 lines of red wines and their 12 attributes.
Decision Trees - A basic decision-making structure, which can be used by a computer to understand and classify data. By asking questions about data items which are fed into them, outputs can be channeled along different branches leading to different results.
Metadata - Data about data, something which might be useful to us.
Predictive modelling - The usage of data to predict the outcome. To do so, a large number of simulated events are used to determine the variables most likely to produce our desired outcome.
Random Forest - It's a method of statistical analysis which involves using outputs of a large number of decision trees and analyzing their outcomes together to provide a better and more understandable classification of data.
Standard Deviation - A common calculation in data science used to measure the difference from a specific result compared to its average.
Testing Data - A part of original data which is used to test the algorithm.
Training Data - A part of original data which is used to train the algorithm.

Data Mining goals:

The plan is to create an algorithm which is best suitable to predict the quality of red wine. To do this, we will train our model from a part of the original data. Ideally, we would like to create a model that is at least 92% accurate. To not be stuck in our training set, we can create training/test data randomly every time by shuffling the dataset and then training/testing models with different data again.
We would also like to find out which variables affect the outcome of wine quality more or less. Correlation between columns will be found.

Task 3:

Gathering Data:

To do our analysis we need at least 1000 rows of data and at least 10 different variable columns. Currently our downloaded dataset contains 1600 rows with 11 variables and outcome (wine quality). We are lucky to have all of our dataset's column values numerical, because it makes it easier for us to train models/find correlations between them. We have already downloaded the file from kaggle and verified its relevance and quality. Currently our dataset has 12 columns (and ID): fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and the outcome - quality. All of them are upon examination relevant to our outcome, however our research and analysis will see if that really is the case. In training our models, we can for example, take all qualities above 7 to be good and beyond bad, and predict if a wine is "good".

Describing, Exploring, Verifying Data:

The description and meaning of each column:

fixed acidity - most acids involved with wine or fixed or nonvolatile (do not evaporate readily). min value 4.6, max 15.9

volatile acidity - amount of acetic acid in wine, too high levels lead to unpleasant, vinegar taste. min value 0.12, max 1.58

citric acid - small quantities of citric acid can add freshness and flavour to wine. min 0.09, max 1

residual sugar - the amount of sugar remaining after fermentation stops. min 0.9, max 15.5

chlorides - the amount of chloride salts in wine. min 0.07, max 0.61

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 wine oxidation. min 7, max 72

total sulfur dioxide - the amount of free and bound forms of SO2, in low concentrations SO2 is almost undetectable in wine, but at free SO2 concentrations over 50ppm, SO2 will impact the smell and taste of wine. min 22, max 289

density - Density is most impacted by the alcohol percent and sugar content. min 0.99, max 1 (a bit pointless it seems)

pH - describes acidity of wine, usually is around 3-4. min 2.74, max 4.01

sulphates - a wine additive which can contribute to SO2 levels. Acts as an antimicrobial and antioxidant. min 0.33, max 2.

alcohol - The percent of alcohol in wine. min 8.4, max 14.9

quality - an output based on sensory data, between 0 and 10. min 3, max 8

The data seems to be usable, although there might be some problems regarding the scale of variables. In some cases, we do not have which scale these are in, which is bad, because we can not predict based on chemical knowledge then. However, some of them do and that is enough for us.

Task 4:

The plan:

1) We are all going to separately analyze data and find correlation and train models, but when we are doing model testing, we will always use randomly shuffled data for train/test sets, to not get stuck in one part of data.
   This part will be individual and takes a different amount of time depending on how fast a person is, every step will take 3-5h. Everything is done in notebook with pandas, numpy, math imports, if any ideas pop up, more can be added.
   This means some of us will:
   1) Find null-correlation by randomly shuffling all columns
   2) Find significance between columns
   3) Modify dataset's column values (for example, instead of quality 1-10 there will be "good - 1" and "bad - 0"
   4) Train different models (like KNN, RF etc)
   5) Find out the combination of settings to get the best model
   6) Create visual outcomes (graphs) between variables, which can be explained easily

2) Then we will report our findings and collaborate on what we think was impactful, what could be changed and how we could together achieve the best result. We will work on our unified notebook and produce the best possible outcome. This part will take time based on how well we do individually and how our outcomes fit. Creating some unified illustrating graphics will also be a priority. Time spent is about 4h, if things go well, more if there are things to discuss. With previous task together, we will all get at least 30h individual work