Student Number: <u>20053722</u>   Name: <u>Bryan Hoang</u>

1. (2 marks) Your first task is to keep a diary, for one of the days of this week, of all the times and places that you notice that data is being collected about you as you live your life. For example, if you buy something online then data about your credit card is obviously collected, but so is data about what kinds of things you buy.

   Try and capture as many different sources as you can, remembering that some data capture is hidden from you. Hand in a list of all of the sources you notice, with brief comments on anything that seems interesting or surprising.

   Answer:

Table 1: Sources of data collected about me on September 13, 2021

Source of Data	Comments
Doing my daily COVID-19 Assessment on the SeQure App in the morning	N/A
Visiting websites with some sort of analytics tracking (e.g., Google Analytics) throughout the day	I usually browse with an ad blocker and cookie blocker to avoid getting tracked as easily. But I know that websites can make use of Browser Fingerprinting to access browsing history and my computer's hardware properties, so I'm always being to be cautious with my privacy on the web as possible.
Google Maps tracking my position while navigating to the appointment in the morning	The app tends to prompt me to review any locations I've visited befores
Filling out some medical forms for a dental related appointment in the morning	As opposed to the above source, I was willing to consent to their collection of my data!
Purchases I made with my credit card at the Dental appointment and in online stores	N/A
Interacting with my Amazon Echo device lets Amazon track usage statistics throughout the day	Privacy concerns around Amazon Echo devices make me wary about how much data the device is sending to Amazon
Signing my name on an attendance sheet in class at around ~14:35	N/A
Listening to music from Spotify throughout the day	Spotify does a yearly summary of people's listening habits (i.e., Spotify 2021 Wrapped)

Student Number: <u>20053722</u>   Name: <u>Bryan Hoang</u>

It's a bit of a stretch, but you could say that this list I'm creating for the exercise question is a source of data being collected about me.	N/A

2. (2 marks) The second task is to use the datasets to get familiar with KNIME. Hand in a brief description of the properties you extracted from the data.

   For example, you might look at the distributions of the attributes, properties such as means and standard deviations, and what the difference between the minimum and maximum values for each attribute are.

   Try to plot some values, including correlations between pairs of attributes.

   Play around with KNIME properties. For example, you could tweak some of the example workflows to do simple predictions.

   Some of these properties could also be investigated using a tool such as Excel. While that is acceptable, you will need to become familiar with KNIME, and the sooner the better, so do everything you can in the KNIME environment.

   Answer:

   For reference, here's a list of the column numbers mapped to the properties described by the data found in the `winedetails.txt` file:

   (1) Kind of wine   (8) Flavanoids

   (2) Alcohol content   (9) Nonflavanoid phenols

   (3) Malic acid   (10) Proanthocyanins

   (4) Ash   (11) Color intensity

   (5) Alcalinity of ash   (12) Hue

   (6) Magnesium   (13) OD280/OD315 of diluted wines

   (7) Total phenols   (14) Proline

Student Number: <u>20053722</u>   Name: <u>Bryan Hoang</u>

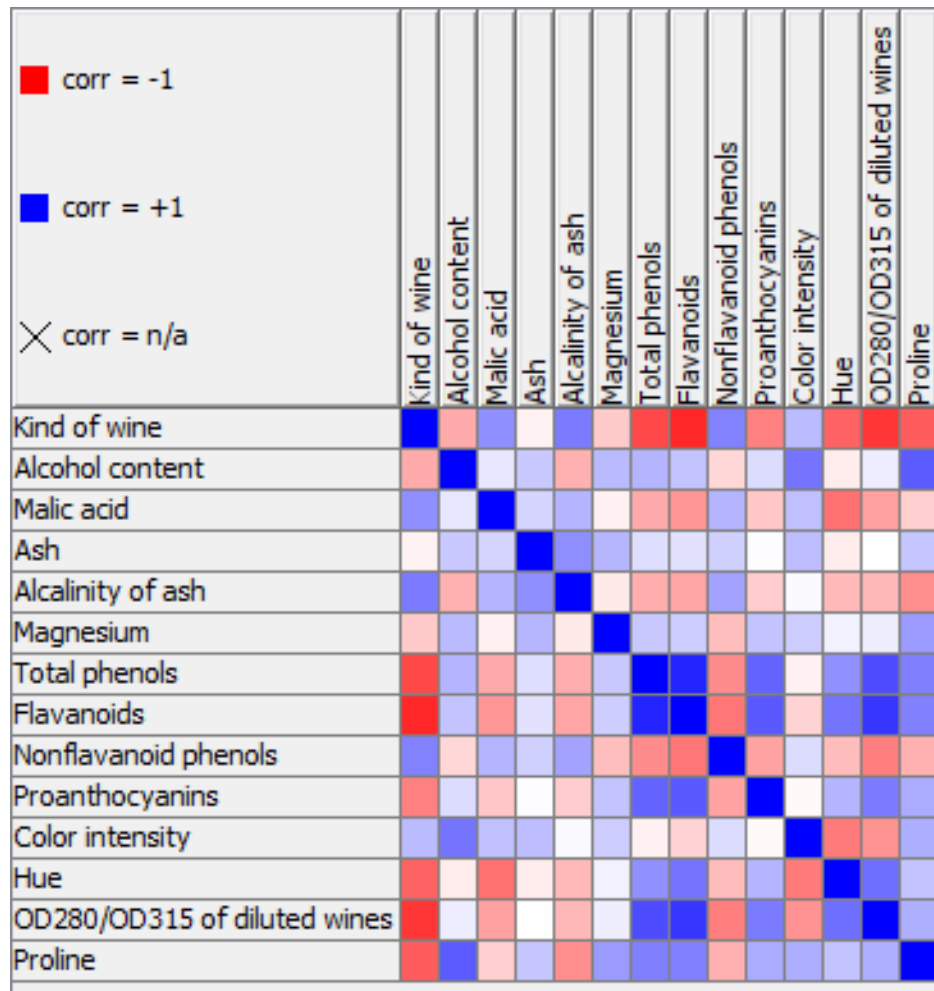To begin analysis, let's observe the correlation matrix of the attributes, as seen in Figure 1.

Figure 1: The linear correlation matrix of the wine data

Notable correlations include:

- the positive correlation between the wine's total phenols and the wine's flavanoids

- the negative correlation between the wine's type (technically a coincidence based on the possible arbitrary numbering) and the wine's total phenols and flavanoids

Student Number: 20053722   Name: Bryan Hoang

Based on the previous observations, let's visually confirm the correlations by observing the line plots of the mentioned attributes previously, as seen in Figure 2.
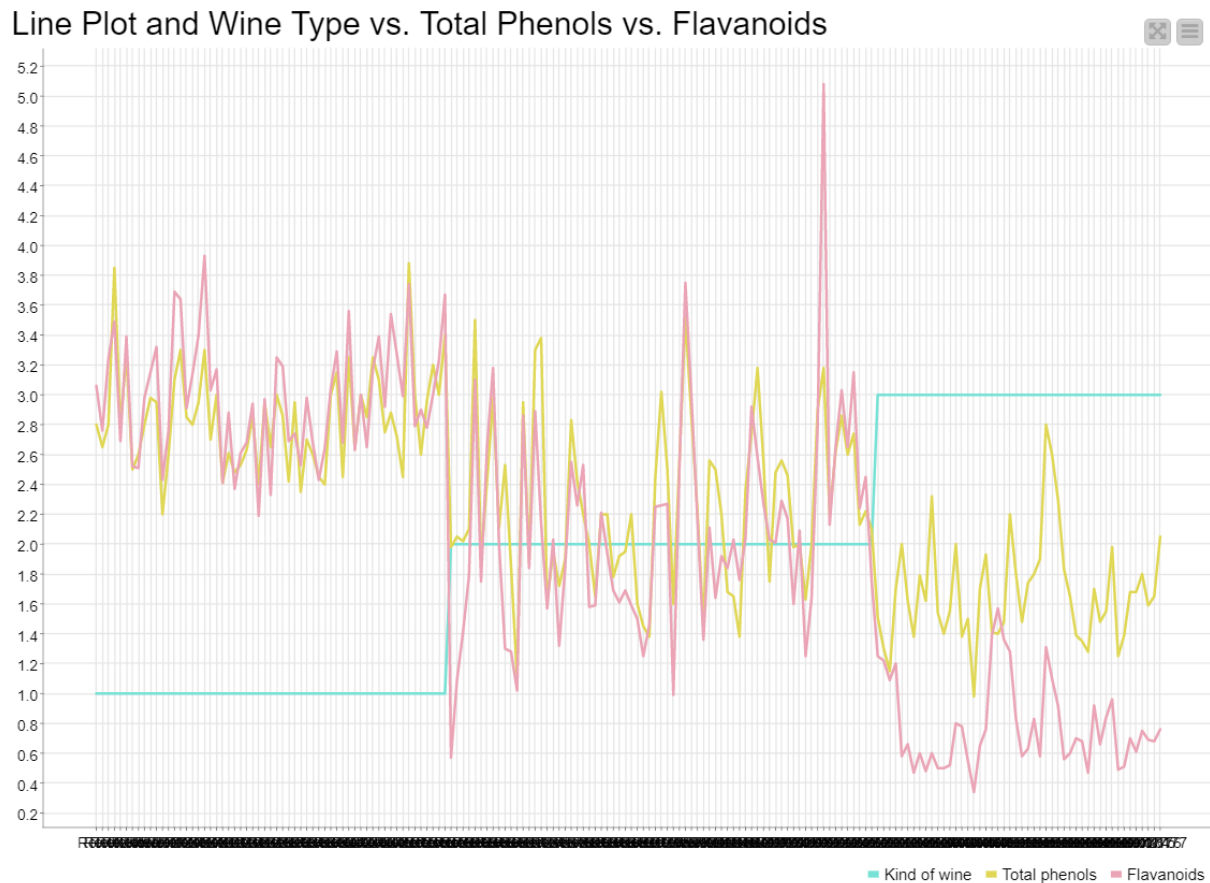
Figure 2: The line plots of the Total phenols, Flavanoids, and the wine's type from the wine data

It is clear that the correlation between the wine's total phenols and the wine's flavanoids is positive and that the correlation between the wine's type and the wine's total phenols/flavanoids is negative.

Using the Statistics Node provided by KNIME, we can also generate a table of all the means, standard deviations, and min/max values for each attribute, as seen in Figure 3.

Student Number: <u>20053722</u>   Name: <u>Bryan Hoang</u>

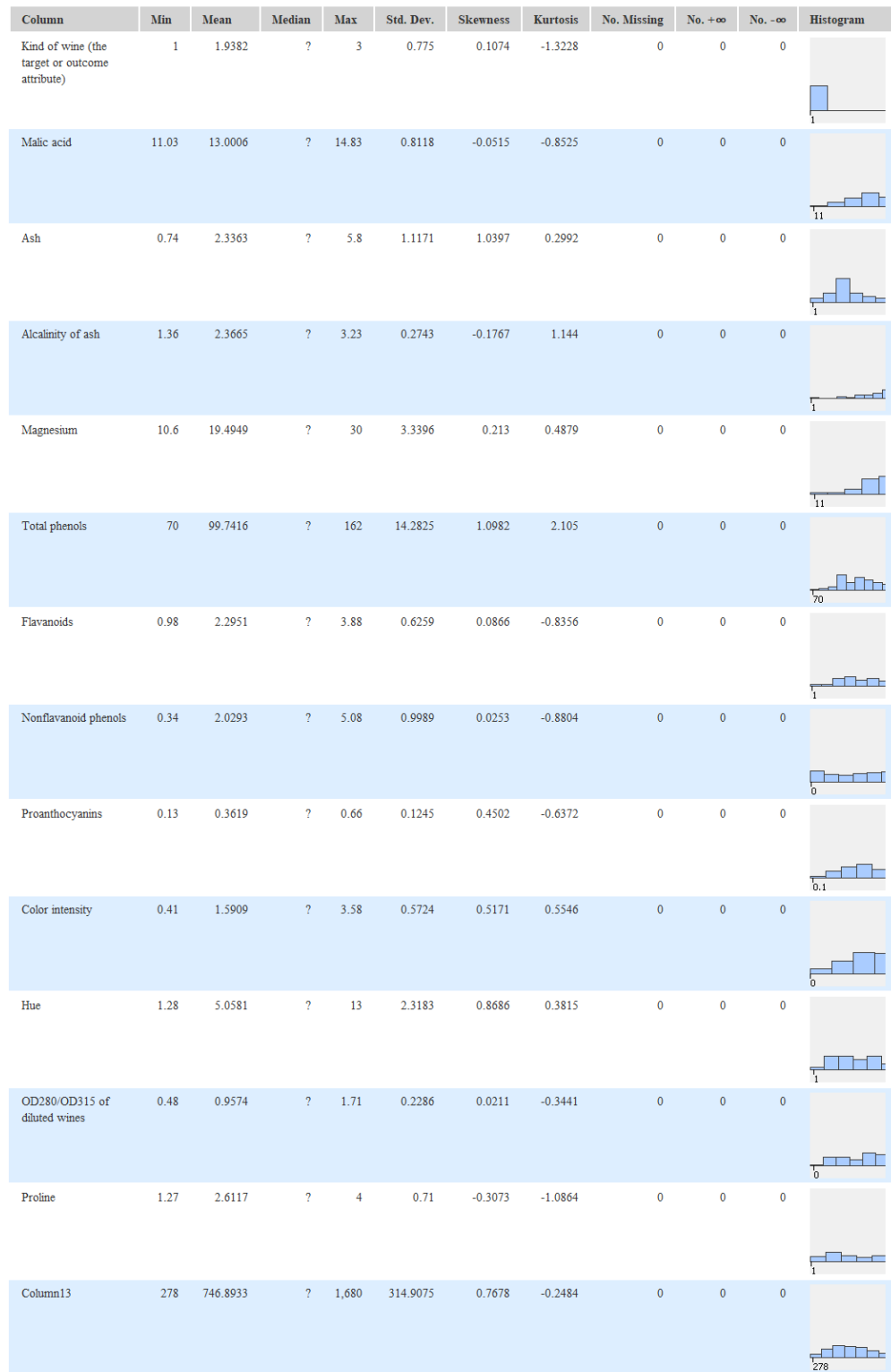| Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
Kind of wine (the target or outcome attribute)	1	1.9382	?	3	0.775	0.1074	-1.3228	0	0	0	
Malic acid	11.03	13.0006	?	14.83	0.8118	-0.0515	-0.8525	0	0	0	
Ash	0.74	2.3363	?	5.8	1.1171	1.0397	0.2992	0	0	0	
Alcalinity of ash	1.36	2.3665	?	3.23	0.2743	-0.1767	1.144	0	0	0	
Magnesium	10.6	19.4949	?	30	3.3396	0.213	0.4879	0	0	0	
Total phenols	70	99.7416	?	162	14.2825	1.0982	2.105	0	0	0	
Flavanoids	0.98	2.2951	?	3.88	0.6259	0.0866	-0.8356	0	0	0	
Nonflavanoid phenols	0.34	2.0293	?	5.08	0.9989	0.0253	-0.8804	0	0	0	
Proanthocyanins	0.13	0.3619	?	0.66	0.1245	0.4502	-0.6372	0	0	0	
Color intensity	0.41	1.5909	?	3.58	0.5724	0.5171	0.5546	0	0	0	
Hue	1.28	5.0581	?	13	2.3183	0.8686	0.3815	0	0	0	
OD280/OD315 of diluted wines	0.48	0.9574	?	1.71	0.2286	0.0211	-0.3441	0	0	0	
Proline	1.27	2.6117	?	4	0.71	-0.3073	-1.0864	0	0	0	
Column13	278	746.8933	?	1,680	314.9075	0.7678	-0.2484	0	0	0	

Figure 3: General statistics of the wine data