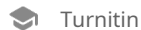


# Applied DS.docx



## Document Details

### Submission ID

trn:oid::27005:73770955

### Submission Date

Dec 11, 2024, 2:05 PM UTC

### Download Date

Dec 11, 2024, 2:06 PM UTC

### File Name

Applied\_DS.docx

### File Size

2.3 MB

14 Pages

1,467 Words

7,986 Characters



# 0% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

**Caution: Review required.**

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

## Detection Groups

- 
**1 AI-generated only 0%**  
 Likely AI-generated text from a large-language model.
- 
**2 AI-generated text that was AI-paraphrased 0%**  
 Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

### Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

## Frequently Asked Questions

### How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (\*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

### What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.



# **CLUSTERING AND REGRESSION ANALYSIS OF ALCOHOL CONTENT: INSIGHTS AND VISUALIZATIONS**

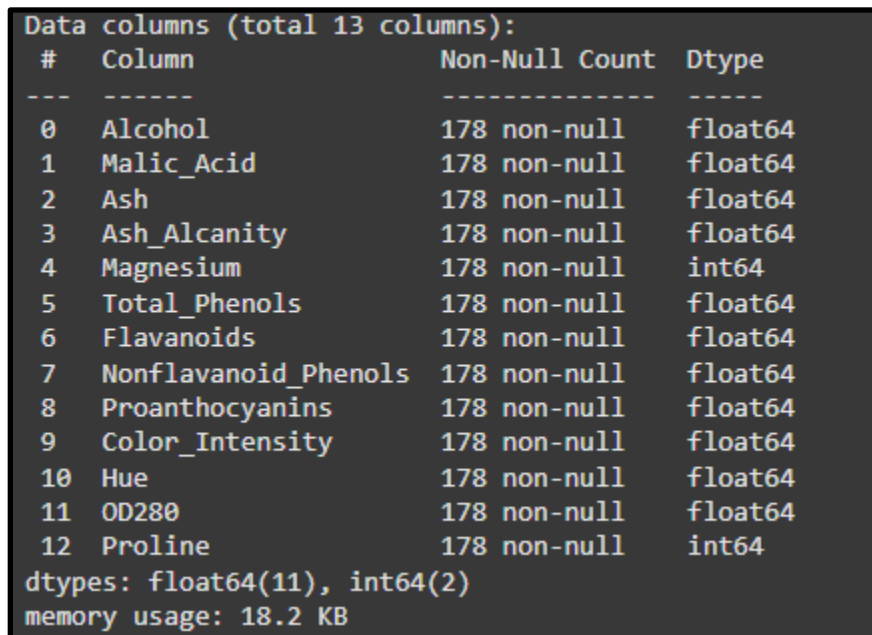
## Table of Contents

1. Introduction:.....	3
2. Data description: .....	3
2.1 Discussion: .....	5
3. Visualization: .....	5
3.1 Discussion: .....	9
4. Clustering: .....	10
5. Linear regression model:.....	11
6. Conclusion: .....	12
7. Bibliography: .....	14

## 1. Introduction:

This paper focuses on clustering and regression analysis on the data set consisting of 13 numerical variables related to alcohol content. Using historical plots, line plots, an elbow plot, and correlation matrices, the report determines patterns of the data, the relationship between variables, and clustering behavior. Such predictions and assessments are done using performance metrics regarding feature importance and determination of the alcohol content. The use of scatter and density plots make the results more understandable while the clustering reveals coherent relations. As depicted in this study, the research findings being employed in an array of statistical and machine learning methodologies give insights regarding relevant alcohol-related data attributes and estimate the capability for identification of high predictive value for diverse characteristics.

## 2. Data description:



```
Data columns (total 13 columns):
#   Column                               Non-Null Count  Dtype
---  -
0   Alcohol                               178 non-null    float64
1   Malic_Acid                            178 non-null    float64
2   Ash                                    178 non-null    float64
3   Ash_Alcanity                           178 non-null    float64
4   Magnesium                             178 non-null    int64
5   Total_Phenols                          178 non-null    float64
6   Flavonoids                             178 non-null    float64
7   Nonflavanoid_Phenols                   178 non-null    float64
8   Proanthocyanins                        178 non-null    float64
9   Color_Intensity                        178 non-null    float64
10  Hue                                    178 non-null    float64
11  OD280                                  178 non-null    float64
12  Proline                                178 non-null    int64
dtypes: float64(11), int64(2)
memory usage: 18.2 KB
```

**Figure 1: Data information**

Figure 1: Dataset attributes which shows that this database contains 13 columns and 178 instances with no missing values. It is composed mostly of data of type float64 with 2 int64 data columns. Concerning the memory usage, the dataset takes 18.2 KB that proves that the data structure is quite

compact. This summary is used to identify which techniques to use when analyzing the data with respect to the types and their sizes.

	Alcohol	Malic_Acid	Ash	Ash_Alcanity	Magnesium \
count	178.000000	178.000000	178.000000	178.000000	178.000000
mean	13.000618	2.336348	2.366517	19.494944	99.741573
std	0.811827	1.117146	0.274344	3.339564	14.282484
min	11.030000	0.740000	1.360000	10.600000	70.000000
25%	12.362500	1.602500	2.210000	17.200000	88.000000
50%	13.050000	1.865000	2.360000	19.500000	98.000000
75%	13.677500	3.082500	2.557500	21.500000	107.000000
max	14.830000	5.800000	3.230000	30.000000	162.000000

	Total Phenols	Flavanoids	Nonflavanoid Phenols	Proanthocyanins \
count	178.000000	178.000000	178.000000	178.000000
mean	2.295112	2.029270	0.361854	1.590899
std	0.625851	0.998859	0.124453	0.572359
min	0.980000	0.340000	0.130000	0.410000
25%	1.742500	1.205000	0.270000	1.250000
50%	2.355000	2.135000	0.340000	1.555000
75%	2.800000	2.875000	0.437500	1.950000
max	3.880000	5.080000	0.660000	3.580000

	Color_Intensity	Hue	OD280	Proline
count	178.000000	178.000000	178.000000	178.000000
mean	5.058090	0.957449	2.611685	746.893258
std	2.318286	0.228572	0.709990	314.907474
min	1.280000	0.480000	1.270000	278.000000
25%	3.220000	0.782500	1.937500	500.500000
50%	4.690000	0.965000	2.780000	673.500000
75%	6.200000	1.120000	3.170000	985.000000
max	13.000000	1.710000	4.000000	1680.000000

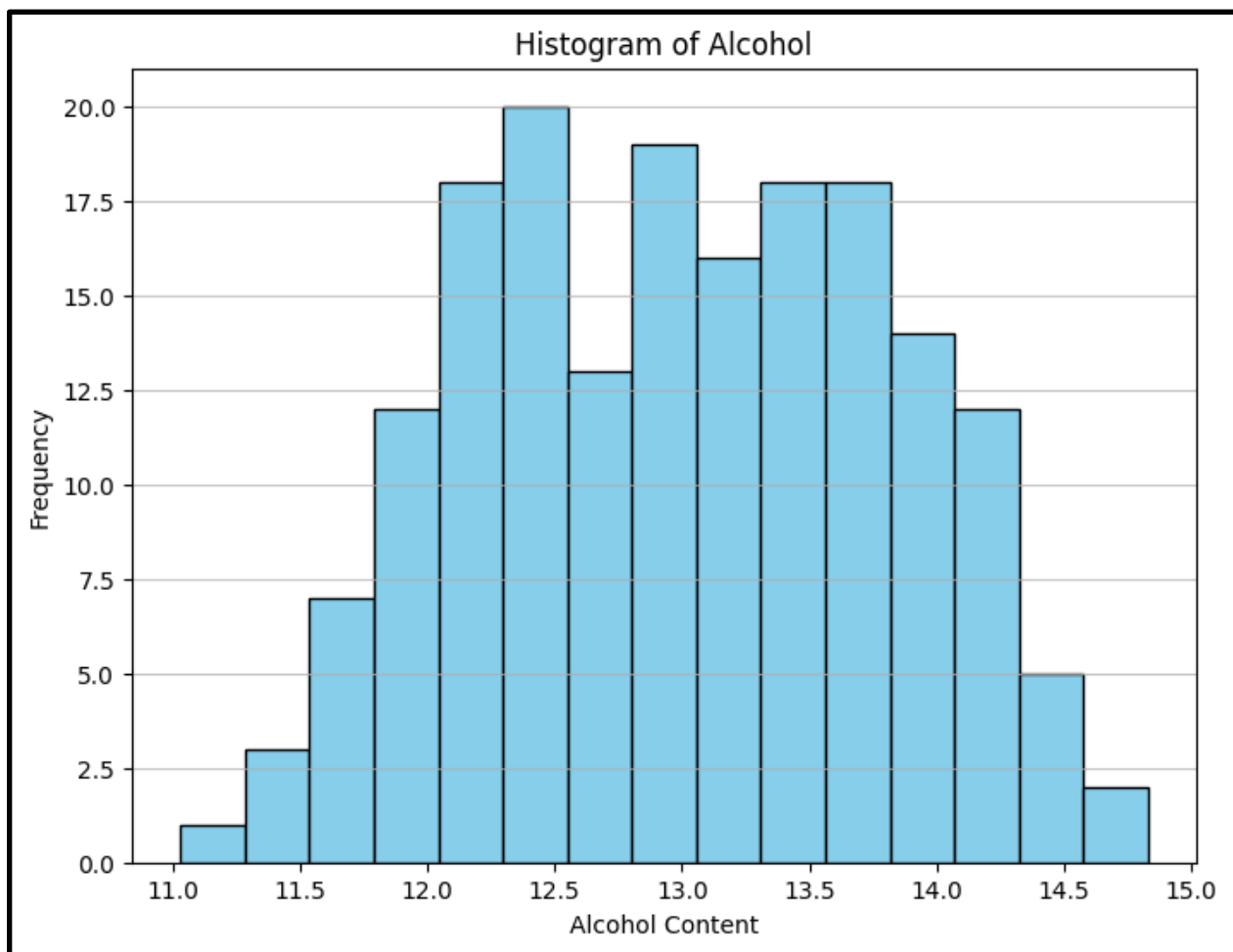
**Figure 2: Summary statistic**

In Figure 2, descriptive statistics of each of the 13 numerical features of the dataset in terms of count, mean, standard deviation, percentiles and maximum values is provided. It also provides information on mean, dispersion and spread of values regarding each attribute. For example in BMI mean is equal to 13.00 while median is equal to 13.05 meaning the data has a very low level of skewness.

## 2.1 Discussion:

In figures 1 and 2 a good insight and overview of the data set that has been used is received. The last two columns of figure 1 supports the adequacy of the dataset where no data are absent and a less sparse data structure. Table 2 provides nominal details about the numerical characteristics, which facilitate the identification of centers, spreads, and gaps, and determine the ensuing analysis.

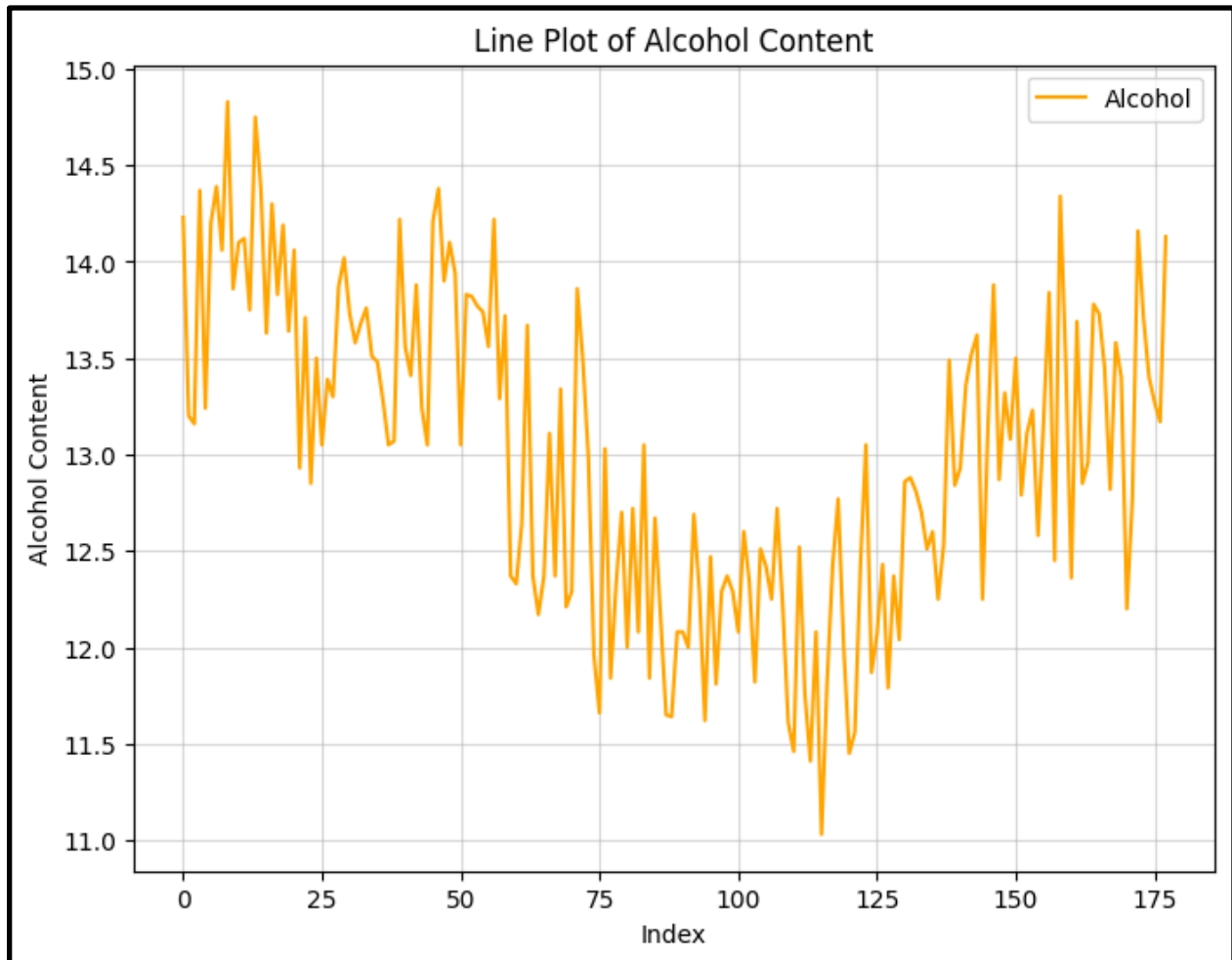
## 3. Visualization:



**Figure 3: Histogram Of Alcohol**

What is more, the histogram of alcohol content presented in figure 3 looks like a normal curve with one peak in the middle of the scale around 12.5. As for the majority of data points, they point to concentration situated between 11.5 and 14.5 alcohol content. A few are hardly located

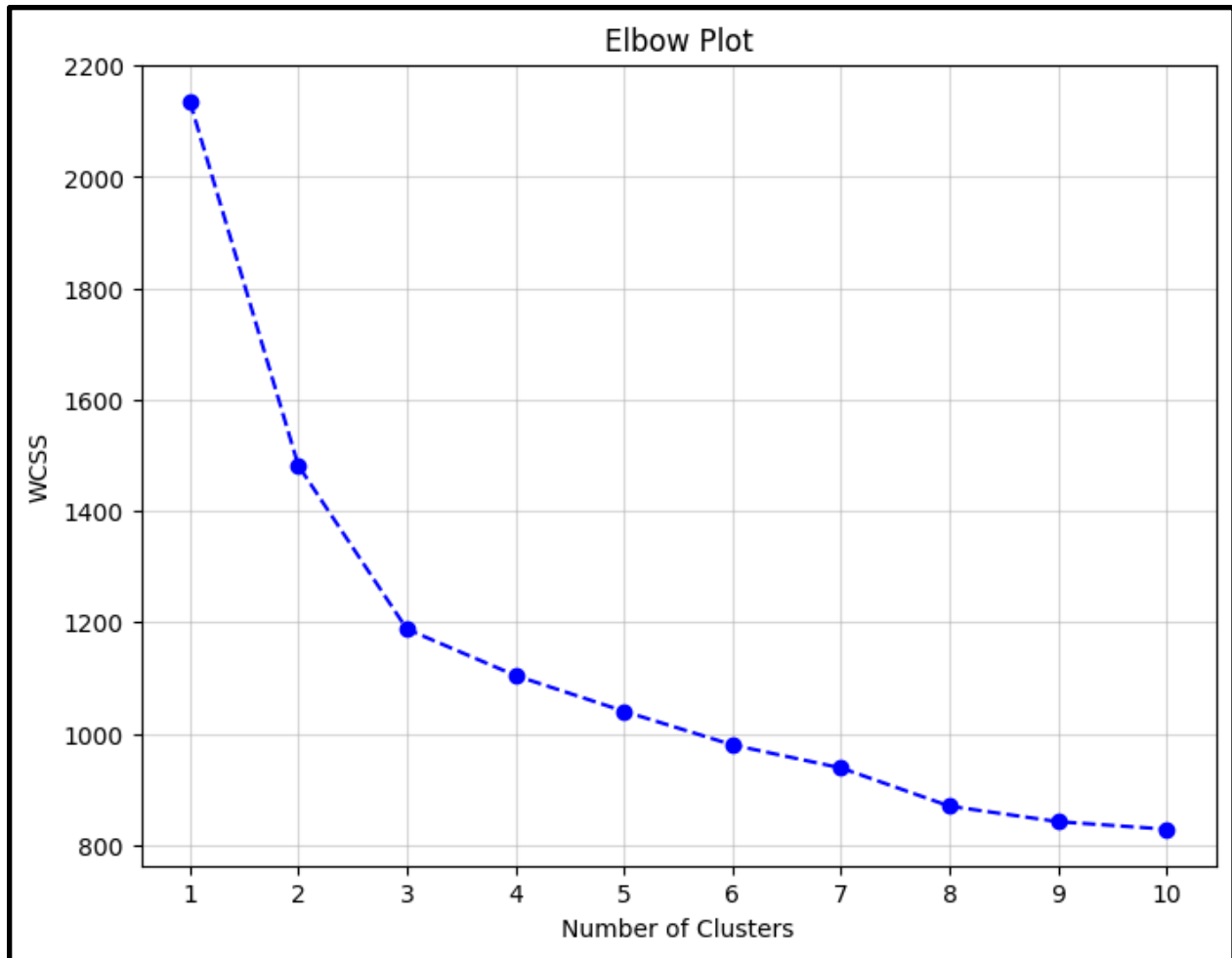
at points of 11 and many are slightly located at points of 14.5 to illustrate that the dataset has some higher and lower alcohol percentage.



**Figure 4: Line Plot of Alcohol Content**

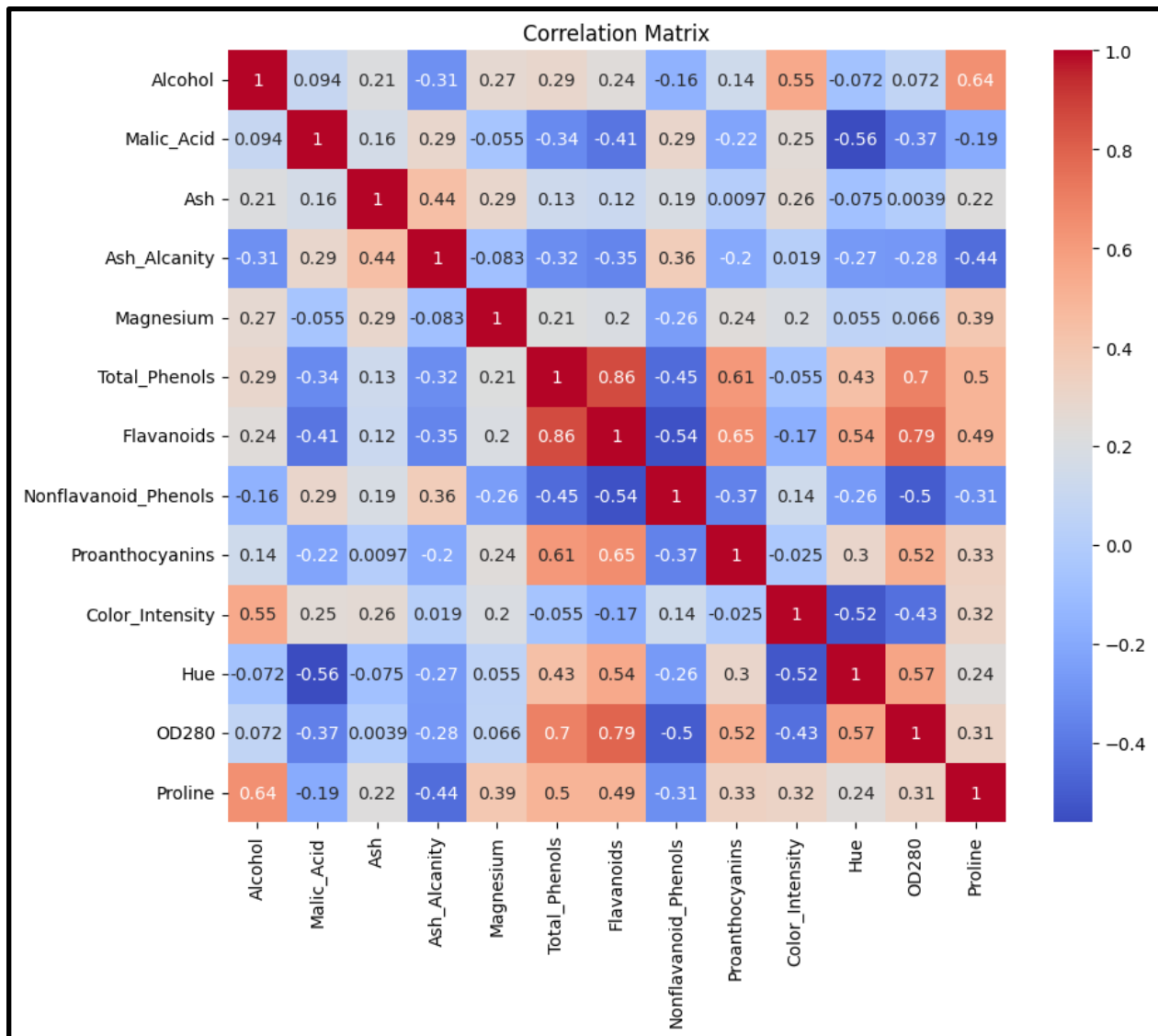
A line plot of alcohol content is shown in Figure 4 to minimize trends or patterns while showing that there are oscillations around the average value. The plot focuses on variation found in the data and helps determine if there are outliers or if the data contains anomalies. This kind of chart is useful in presenting the distribution of the data and in identifying deviant conferences in alcohol content.





**Figure 5: Elbow Plot**

The elbow plot used in the current study, indicating the number of clusters appropriate for the K-means analysis is shown in figure 5. The number of clusters is on the x-axis and the sum of squares within clusters on the y-axis stands for 'WCSS'. WCSS is an inverse function of the number of clusters and 'elbow' point chosen at 3 clusters represents a high balance between clustering efficiency and model simplicity.



**Figure 6: Correlation Matrix**

Finally, Figure 6 illustrates the correlation matrix with regard to numerical features, which demonstrated pairwise proportions correlation. It is both balanced and centred along the diagonal wherein diagonal values are set at one due to auto-correlation. Colours identify correlation intensity and sign, while red means positiveness, and blue means negativeness. These are presented as strong positive co-efficient of Flavonoids and Total Phenols and strong negative co-efficient of Nonflavanoid Phenols and Proanthocyanidins, which helps in feature selection.

### 3.1 Discussion:

Overall, Figures 3, 4, 5, and 6 summarize the main aspects of the dataset considered in the present paper. From the histogram in figure 3, it can be deduced the central tendency of alcohol content and variability, it is observed that it followed a bell shaped distribution since most products clustered at middle range of alcohol content. The line plot of sample by sample variability given in Figure 4 is also useful to identify whether there are any obvious trends in the variability. As for the elbow plot in Figure 5, the most appropriate K-means clustering is decided to be three since it is the actual number for the formation of clusters. In the feature space correlation matrix shown in figure 6, it is easier to discover strong correlations between features such as Flavanoids with Total Phenols to help in feature engineering and for guidance in clustering.

## 4. Clustering:

Cluster centers:

```
[[ -0.35315619 -0.46687385  0.23139163 -0.64403942 -0.16593847 -0.02800099
   0.06980726 -0.11213746 -0.92651429  0.47458392  0.27012353 -0.82867165]
 [  0.87154706  0.18689833  0.52436746 -0.07547277 -0.97933029 -1.21524764
   0.72606354 -0.77970639  0.94153874 -1.16478865 -1.29241163 -0.40708796]
 [ -0.34733103  0.25430244 -0.58207906  0.59715068  0.86333733  0.92176358
  -0.59533423  0.67056519  0.08288917  0.46200513  0.72820041  0.99745567]]
```



**Figure 7: Clustering**

Analysis of the clustering is presented in additional Figures 7, which also use scatter and density plots to visualize the results. The diagonal shows distribution of density plots of individual features

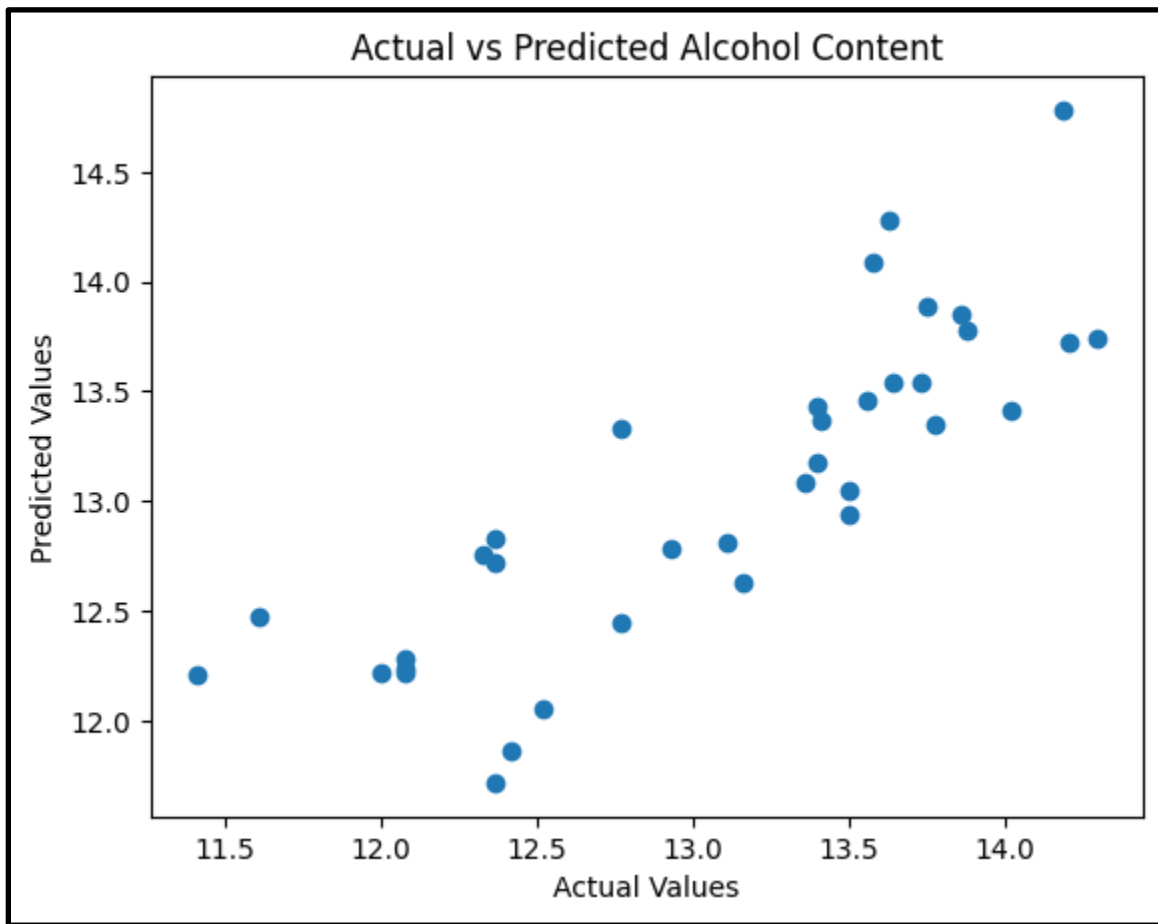
which represent the distribution and spread of the features. Off-diagonal sections contain histograms that show feature distribution and their correlation distributions for pairs of attributes. Locations of the cluster centers are shown at top where each point represents the average of the features for a cluster. They assist in defining the different formation of the clustered groups to which these centers belong to. This figure also helps identify the distribution of data, relationship between features, and how well clustering technique isolates data points for better analysis in subsequent stages.

## 5. Linear regression model:

```
Mean Squared Error: 0.18571608706852522
R^2 Score: 0.6889364317843805
Coefficients of the regression model:
Malic_Acid      0.112887
Ash             0.268318
Ash_Alcanity    -0.046854
Magnesium       -0.001801
Total_Phenols   0.091979
Flavanoids      -0.010802
Nonflavanoid_Phenols -0.124330
Proanthocyanins -0.142778
Color_Intensity 0.169749
Hue             0.403695
OD280           0.119319
Proline         0.000990
dtype: float64
```

**Figure 8: Coefficients of the regression model**

In figure 8 below are the coefficients of a regression model that shows the impact of each feature to the target variable. Coefficients are just numbers and can also be positive or negative, this reflects the idea of strength in relationships. It is important to note that positive sign indicates direct relationship or in other words direct impact, negative sign implies an inverse relationship or in other words negative impact. The figure includes performance metrics: The analysis of the prediction error is presented by the Mean Squared Error (MSE) and the measure of the model variability is presented by the R-squared. This plot helps in understanding feature importance and assess the predictive capability of the model and also pinpoints how specific features contribute to the values of regression.



**Figure 9: Actual vs Predicted Alcohol Content**

Using a scatter plot in Figure 9, actual and predicted values of alcohol content are compared. Larger points are more noted near the diagonal line, which suggest the reasonableness of the model in the predictions. Using deviations from the line, it is possible to demonstrate that predictions did not turn out to be as expected. The dispersion around the diagonal gives an idea of variability of prediction which gives an overall idea of the efficiency of the regression equation in predicting the alcohol content of the beers.

## 6. Conclusion:

It is found that this analysis is able to provide relevant insights by using both clustering and regression analysis to the alcohol content data set. For the relationships between features, the correlation matrix and scatter plot could be used and for the structure of clustering, the elbow plot could be conveniently used. Regression results pay more attention to feature importance and

qualitatively confirm the model accuracy using measures such as R-squared and Mean Squared Error. The analysis of the predicted and actual alcohol content also proves the credibility of the models as slight disparities remain sources of improvement. The findings of this report illustrate how concepts used in analytical tools may be applied to a real-world set of data, to facilitate analytical decision making and predictive modeling in similar contexts.

## 7. Bibliography:

Agin-Liebes, G., Nielson, E.M., Zingman, M., Kim, K., Haas, A., Owens, L.T., Rogers, U. and Bogenschutz, M., 2024. Reports of self-compassion and affect regulation in psilocybin-assisted therapy for alcohol use disorder: An interpretive phenomenological analysis. *Psychology of Addictive Behaviors*, 38(1), p.101.

Bell, L., Garnett, C., Qian, T., Perski, O., Williamson, E. and Potts, H.W., 2020. Engagement with a behavior change app for alcohol reduction: data visualization for longitudinal observational study. *Journal of medical Internet research*, 22(12), p.e23369.

Bujalski, M., 2024. Exploring public preferences for alcohol risk communication. *Drugs: Education, Prevention and Policy*, pp.1-10.

Lim, Y.S., Kim, J.S., Choi, J.H., Kim, J.M. and Shim, T.S., 2022. Solvatochromic discrimination of alcoholic solvents by structural colors of polydopamine nanoparticle thin films. *Colloid and Interface Science Communications*, 48, p.100624.