Wine KMC_MD Data | Individual Assignment

Business Understanding

Objective:
The primary goal of this analysis is to optimise the wine deal offerings for the import-export business specialising in bulk wine. The business aims to enhance its profitability by strategically selecting and presenting wine deals in the monthly email newsletters.

The conducted analysis aims to identify patterns and predict the key factors that contribute to the success of wine deals in terms of sales, ultimately proposing a data-driven and adaptive methodology for ongoing improvements to enhance the efficacy of future wine deals.

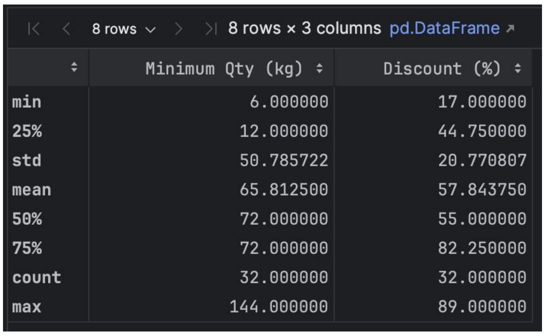
Descriptive Statistics: Data Understanding

Variables:
The dataset includes 32 observations and 8 variables which are:

1. Offer # - int64
2. Customer Last Name - object
3. Campaign - object
4. Varietal - object
5. Minimum Qty (kg) - int64
6. Discount (%) - int64
7. Origin - object
8. Past Peak - bool

Statistical Summary:
Getting the statistical summary of the central tendency, spread, and distribution of the data for the numerical variables "Minimum Qty (kg)" and "Discount (%)".

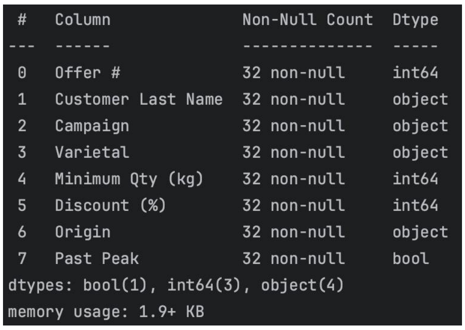
The mean amount of sales is **65.81** and the mean amount of discount proposed by the salesperson is **57.84**.


8 rows
8 rows × 3 columns pd.DataFrame ↗

	Minimum Qty (kg)	Discount (%)
min	6.000000	17.000000
25%	12.000000	44.750000
std	50.785722	20.770807
mean	65.812500	57.843750
50%	72.000000	55.000000
75%	72.000000	82.250000
count	32.000000	32.000000
max	144.000000	89.000000

Table 1

Missing Values:
Upon exploring the data, no missing values were found. The dataset is complete and ready for analysis without any gaps.



```
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Offer #             32 non-null     int64
 1   Customer Last Name  32 non-null     object
 2   Campaign            32 non-null     object
 3   Varietal            32 non-null     object
 4   Minimum Qty (kg)    32 non-null     int64
 5   Discount (%)        32 non-null     int64
 6   Origin              32 non-null     object
 7   Past Peak           32 non-null     bool
dtypes: bool(1), int64(3), object(4)
memory usage: 1.9+ KB
```

Table 2

Outliers:
In examining the numerical variables, "Minimum Qty (kg)" and "Discount (%)" visualising with a boxplot, no apparent univariate outliers were observed. Subsequent verification using the Z-score method confirmed the absence of univariate outliers in these variables.

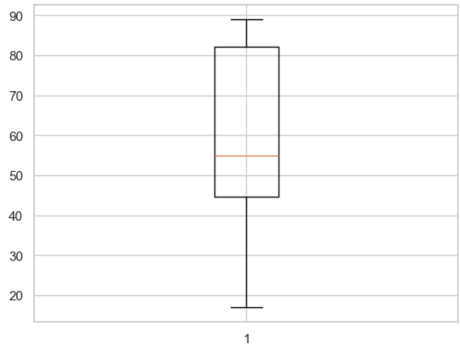


Figure 1

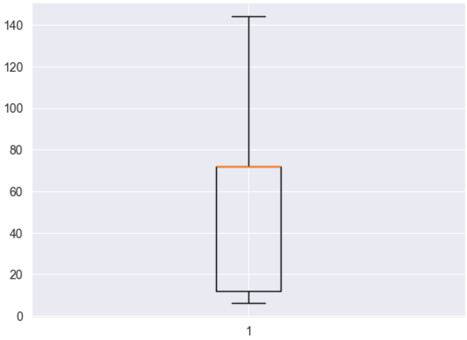


Figure 2

To further scrutinise the data, the Mahalanobis method was employed to identify multivariate outliers. The resultant array was found to be empty, indicating the absence of any multivariate outliers within the dataset. This comprehensive outlier analysis provides assurance regarding the robustness of the data with respect to both univariate and multivariate considerations.

Checking for Normality:
Upon thorough examination of the data's normality using boxplots and qqplots, it's clear that the 'Minimum Qty' variable deviates significantly from a normal distribution. Although the 'Discount %' variable appears somewhat normal based on the qqplot, it doesn't exhibit sufficient normality to be conclusively considered normally distributed. Given the inherent characteristics of the data, the non-normal distribution of these variables is expected and acceptable. It is crucial to note that traditional modelling techniques, which assume normality, may not be suitable for these variables. Therefore, in the further steps selecting a modelling technique that is robust to the non-normality of the data is essential.

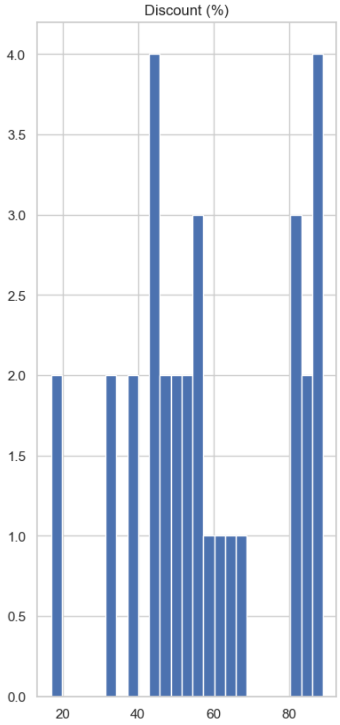


Figure 3

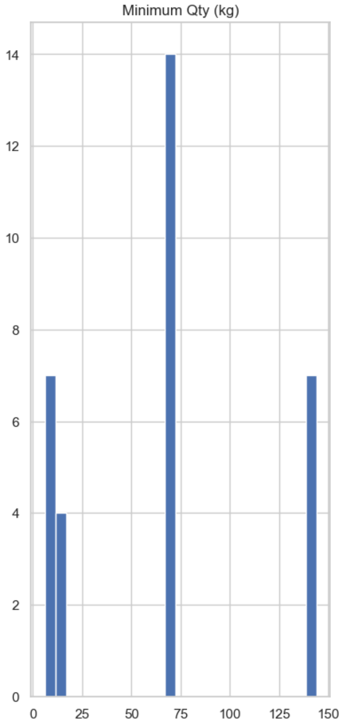


Figure 4

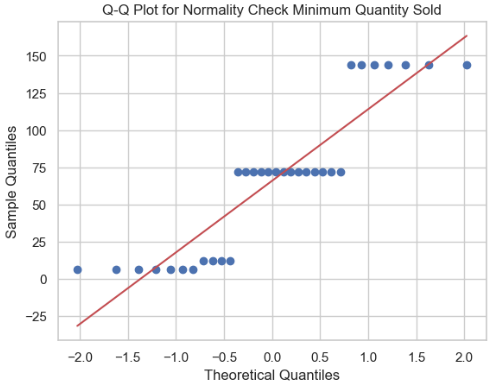


Figure 5

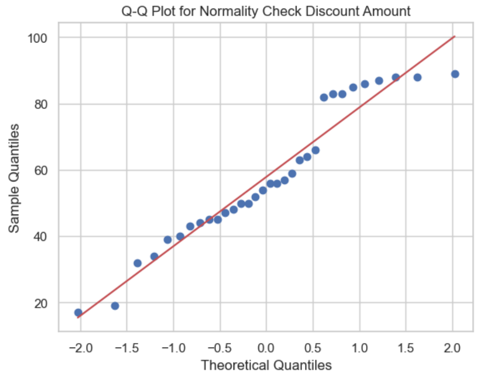
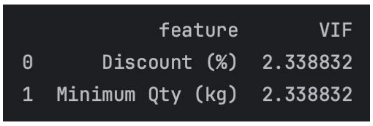


Figure 6

In order to further assess the normality of the distributions for both ‘Minimum Quantity’ and ‘Discount’, a Shapiro-Wilk test was performed. The obtained results indicate a Shapiro-Wilk test statistic of **0.9079** and a corresponding p-value of **0.0001611**. Given that the p-value is below the conventional significance threshold of 0.05, we reject the null hypothesis. Consequently, we can assert that the distributions for Minimum Quantity in kg and Discount do not adhere to a normal distribution.

Checking for Multicollinearity:
After assessing the data, it is evident that there is no significant correlation between the discount offered and the resulting sales. The correlation coefficient between the discount and minimum quantity is **0.0286**, indicating a **weak positive correlation**. This suggests that factors other than discounts play a more influential role in driving sales. The salesperson may benefit from directing his attention toward these other factors, which will be further examined in the report.

The VIF values of approximately **2.33** for both "Discount (%)" and "Minimum Qty (kg)" indicate a moderate level of correlation with other variables, but they do not surpass the commonly recognized threshold of 10, suggesting there are no significant concerns regarding multicollinearity.



```
            feature       VIF
0      Discount (%)  2.338832
1  Minimum Qty (kg)  2.338832
```

Table 3

Checking for Linearity:
After a thorough examination for linearity, as evidenced by a scatter plot analysis, it is apparent that the variables 'Minimum Qty' and 'Discount %' do not exhibit a linear

dependency. Consequently, the assumption of linearity remains unviolated for these numerical variables.

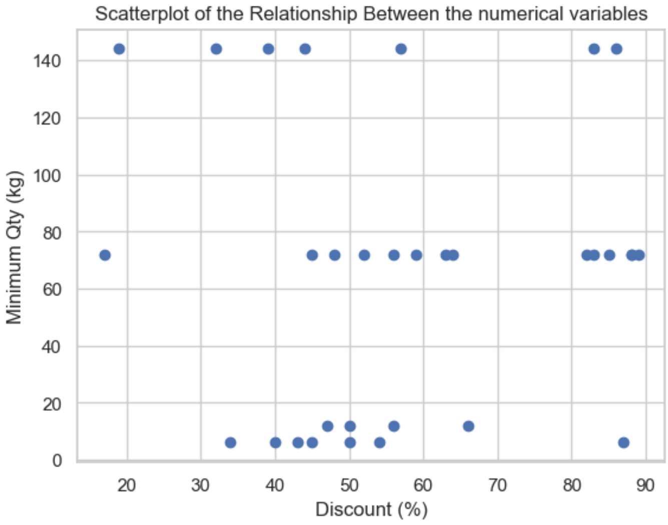


Figure 7

Patterns & Relationships

Past Peak & Varietal:
Examining the boolean variable "past peak" by summing the number of TRUE observations reveals an intriguing observation: only **7** of the wines are classified as **past peak**. In the context of wines, the term "Past Peak" typically denotes a stage in the wine's maturation process where it has reached its optimal ageing potential and may begin to show signs of decline in quality. The fact that only a small fraction of the wines are labelled as "Past Peak" suggests that the majority may still be in a favourable state of development or have not yet reached their optimal maturation point.

Exploring the relationship between the varietal of the wine and its aging status, a notable pattern emerges: Sauvignon wines tend to be predominantly labelled as 'Past Peak,' indicating a preference for aging. In contrast, other types of wines exhibit a higher likelihood of being categorized as 'not past peak,' suggesting a tendency for these varieties to be served in a fresher state rather than aged. This observation hints at distinct aging preferences across different varietals, adding nuance to our understanding of how aging characteristics vary within the spectrum of wine types.

Past Peak & Minimum Quantity:
Upon investigating the 'Past Peak' variable in the dataset, a question arises: Does the ageing status of wines impact sales? Is there a tendency for more aged wines to be purchased?

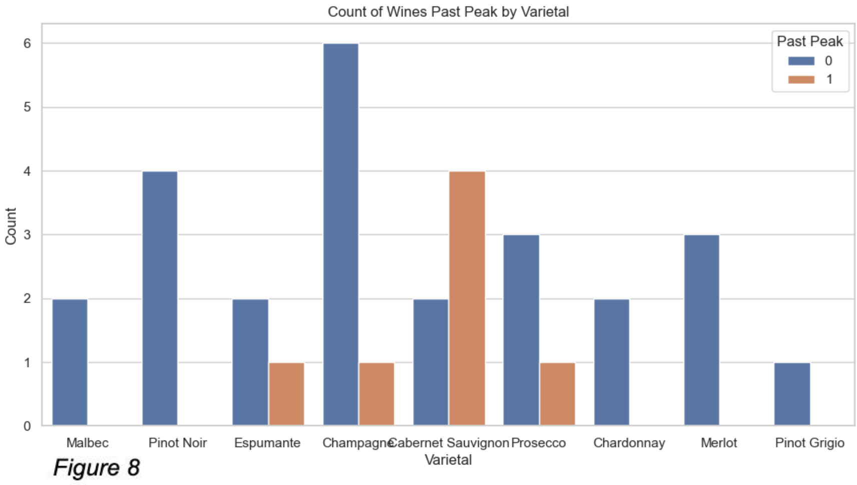


Figure 8

After visualising the relationship with a bar plot, a compelling pattern emerges—wines labelled as 'past peak' indeed exhibit higher sales. Interestingly, despite this trend, the majority of supply made by salesmen are not wines past their peak. Out of 32 observations, only 7 were classified as 'past peak,' yet they demonstrated significant sales figures. This suggests that customers of the salesmen may lean towards a diverse selection that includes a variety of aged wines.

It's essential to note that the observed patterns may benefit from further exploration with a larger dataset. With only 32 observations, the full extent of the relationship between aging status and sales might not be entirely clear. Additional data could provide more robust insights into customer preferences regarding the purchase of past peak wines.

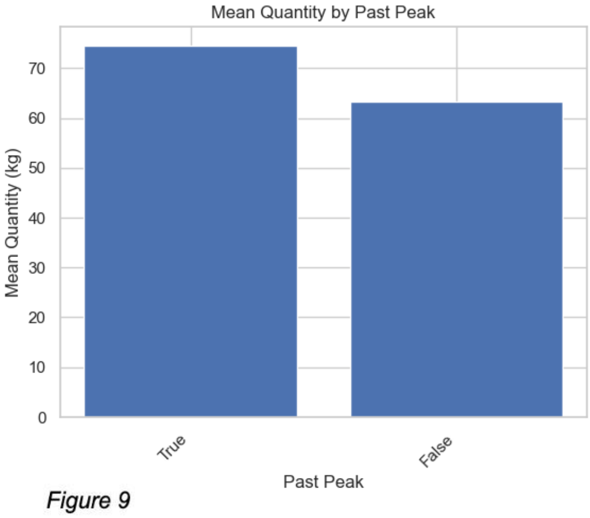
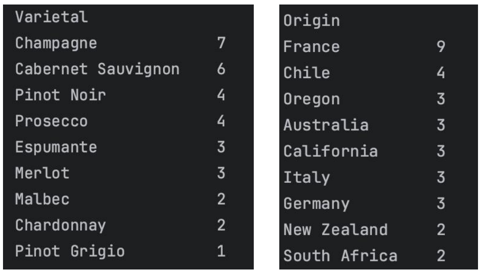


Figure 9

Varietal, Region & Minimum Quantity:

From *Table 4*, it is evident that the most imported wine types are Champagne, Cabernet Sauvignon, and Pinot Noir.

Additionally, the primary countries of origin for the wines are France and Chile. These nations dominate the list of top exporters, indicating a strong presence in the supply that the salesperson offers.



```
Varietal
Champagne             7
Cabernet Sauvignon    6
Pinot Noir            4
Prosecco              4
Espumante             3
Merlot                3
Malbec                2
Chardonnay            2
Pinot Grigio          1
```

```
Origin
France          9
Chile           4
Oregon          3
Australia       3
California      3
Italy           3
Germany         3
New Zealand     2
South Africa    2
```

Table 4

A surprising pattern emerges upon scrutinising the correlation between customers' preferences and the offerings from the salesmen. When plotting the Varietal variable against the sales amounts we can notice that the customers' preferred choice are Chardonnay and Cabernet Sauvignon, surpassing even Champagne in popularity.

Curiously, the salesmen import a relatively limited quantity of Chardonnay wines. Expanding the selection of Chardonnay and Cabernet Sauvignon in their offerings could potentially lead to a considerable boost in sales.

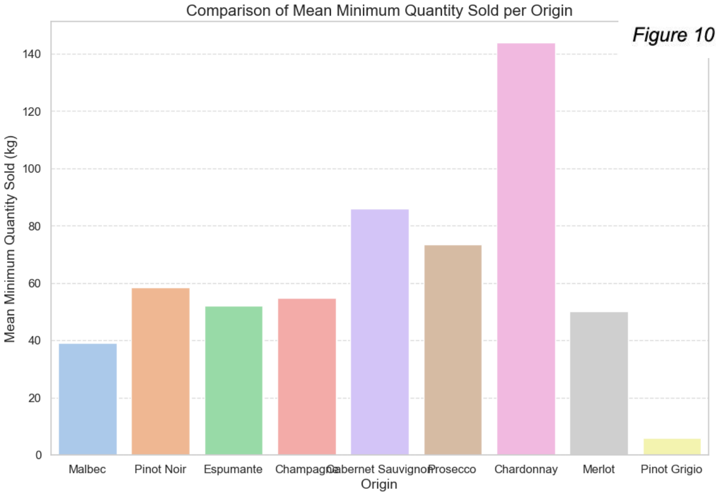


Figure 10

Contrary to a common stereotype that assumes a higher preference for wines from France due to its renowned status, a deviation appears in the sales data. It appears that certain regions surpass France in terms of customer preferences. Notably, New Zealand and Chile stand out with higher sales figures.

In light of this observation, the salesperson might want to contemplate increasing imports from New Zealand while maintaining the current pace for Chilean wines. Additionally, other regions such as Oregon, Australia, South Africa, and Italy also outperform France in sales. This could be indicative of a shift in customer preferences towards exploring new flavours rather than following the traditional choices. Further research is needed to uncover the latent factors contributing to this trend and to inform strategic decisions regarding the import strategies.

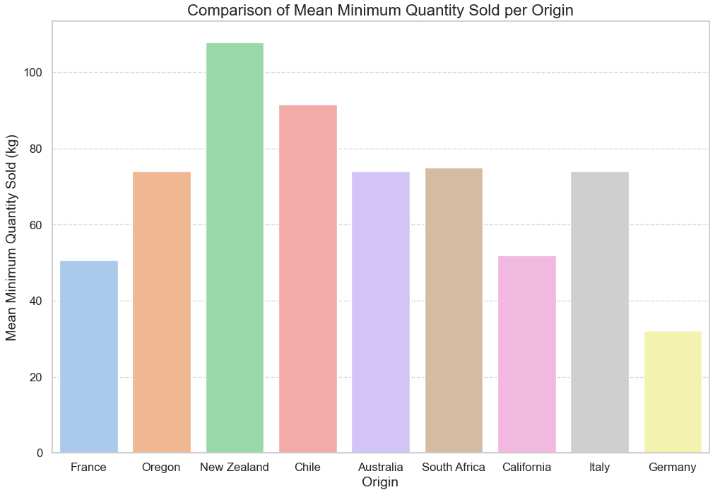


Figure 11

In the graph, the predominant import is Champagne from France. However, the analysis suggests that it may not align with customer preferences. Instead, increasing imports of Chardonnay from Chile and South Africa, along with Cabernet Sauvignon from New Zealand, Italy, and Oregon, could better cater to customer choices and potentially boost overall sales.

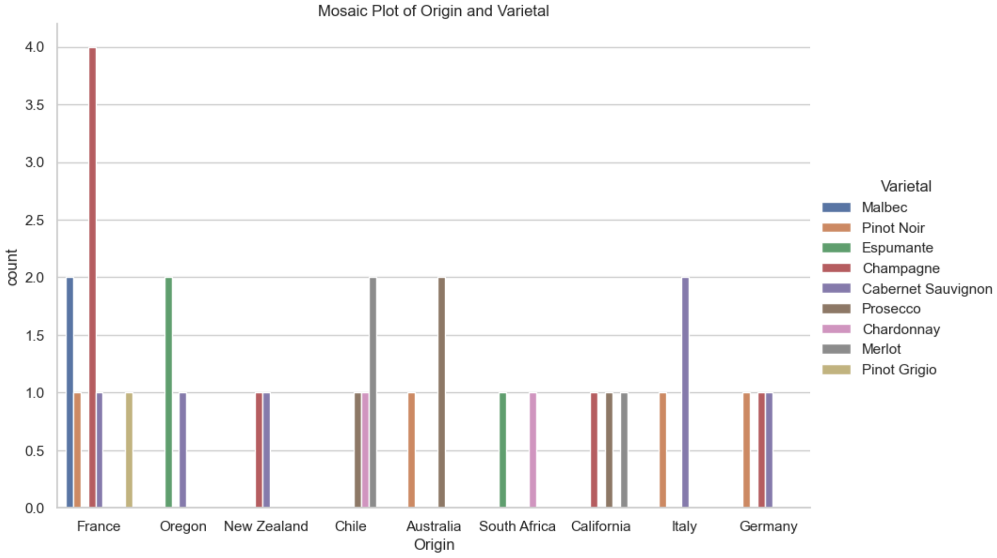


Figure 12

Seasonality | Campaign & Minimum Quantity Sold:
Analysing the correlation between campaign months and sales figures reveals intriguing patterns. Notably, campaigns conducted in February, April, June, and October exhibit a notably higher efficacy compared to those in November and July. However, it's crucial to acknowledge that these observations, while suggestive, are insufficient for definitive conclusions. With a more extensive dataset, we can potentially detect a clearer trend and draw more robust insights into the effectiveness of campaigns during different months.

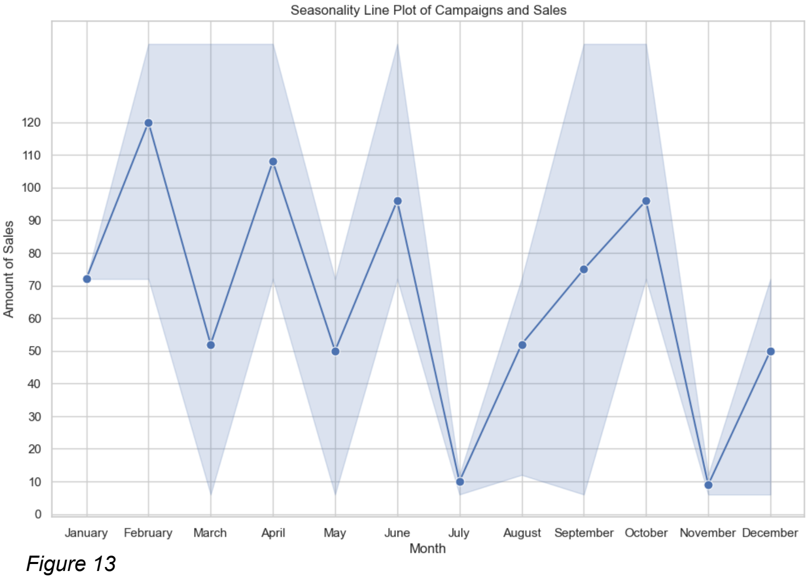
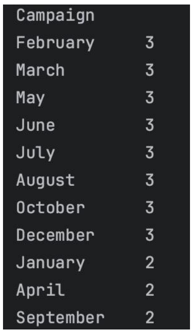


Figure 13



```
Campaign
February     3
March        3
May          3
June         3
July         3
August       3
October      3
December     3
January      2
April        2
September    2
```

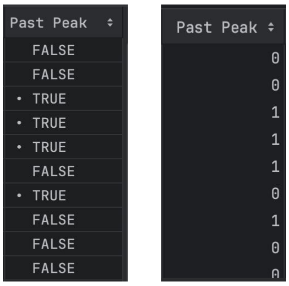
Table 5

Clustering

The selection between Clustering and Principal Component Analysis (PCA) depends on the inherent nature of the data and the overarching goals of the analysis. In the current context, the preference leans towards clustering for a deep exploration of natural groups or patterns within the Wine dataset. The objective is to explore customer preferences regarding specific wine groups, fostering conclusive insights.

Moreover, the decision to opt for clustering is reinforced by the presence of both categorical and numerical variables in the dataset. Unlike PCA, clustering algorithms handle this mixed-type data, making them a more fitting choice for our comprehensive analysis. It's important to mention that PCA, designed for reducing the number of variables/features, is less useful when working with only eight variables, especially when excluding two of them, ‘Offer #’ and ‘Customer Last Name’ from the analysis.
Last but not least, the efficacy of clustering extends to capturing non-linear relationships and intricate patterns within the data. This quality becomes particularly significant given the noted non-linear pattern between ‘Minimum Quantity’ and the associated ‘Discount’. Consequently, as clustering works well with different types of data, fits the size of our dataset, and can reveal complex relationships, it's the better choice for the analysis.

Step 1:
Converting the boolean variable to integers to facilitate its inclusion in the clustering process.



```
Past Peak ÷
  FALSE
  FALSE
• TRUE
• TRUE
• TRUE
  FALSE
• TRUE
  FALSE
  FALSE
  FALSE
```

```
Past Peak ÷
          0
          0
          1
          1
          1
          0
          1
          0
          0
```

Table 6

Step 2:

Developing a Hierarchical model and generating a dendrogram is an initial step to discern the potential number of clusters. In this analysis, a threshold of 1.5 was selected to aid in identifying distinct clusters. While Hierarchical clustering provides valuable insights, recognizing its basic nature, the analysis with the K-means clustering method is also performed. The combination of these methods enhances our understanding and facilitates a more robust exploration of optimal cluster configurations.

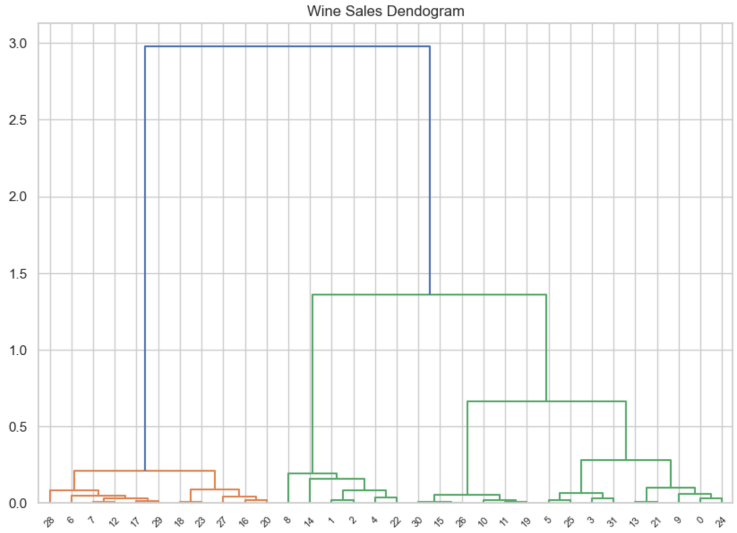


Figure 14

Step 3:

Upon dendrogram analysis, it's evident that the optimal number of clusters is 3, supported by the Elbow method which indicates a significant change at the 3-4 cluster transition. Given the small size of the dataset, I've opted for 3 clusters as the most suitable K amount to ensure a meaningful segmentation of the data.

Step 4:

The K-means analysis was executed, resulting in the formation of the following clusters:

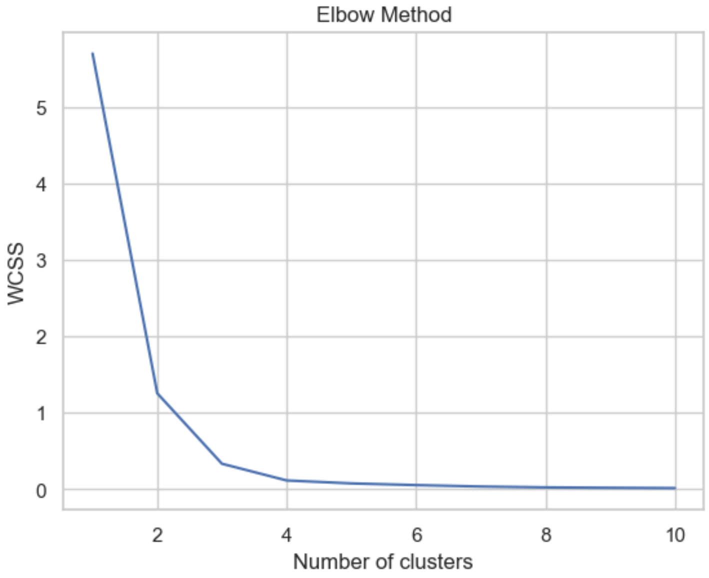


Figure 15

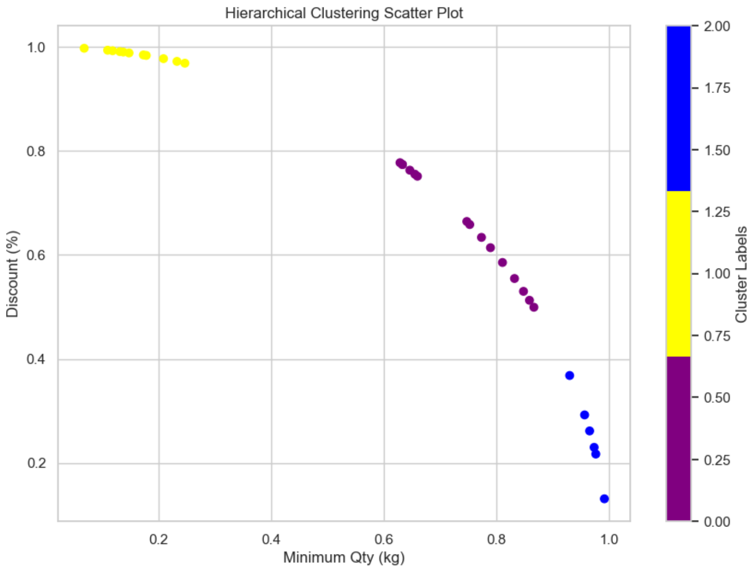
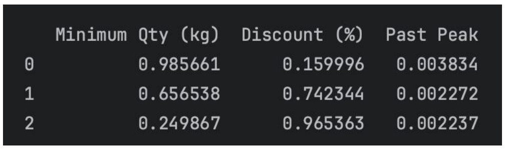


Figure 16

There appear to be distinct variations among the cluster groups. In the initial cluster, customers are making substantial purchases even when the discount amount is not a primary motivating factor, following the earlier indications in the report. In contrast, the second cluster places significance on discounts, yet their purchasing volume does not exceed that of the first cluster so there's need to put emphasis on both factors.

Remarkably, the third cluster, despite being offered the highest discounts, exhibits the lowest purchasing activity. In this segment, there seems to be a reversal of expectations—perhaps these customers are influenced by factors other than discounts and may be seeking additional features or attributes in the email newsletter.

The variable "Past Peak" appears to have negligible influence on the clustering outcomes, as its values are considerably small.



```
   Minimum Qty (kg)  Discount (%)  Past Peak
0          0.985661      0.159996   0.003834
1          0.656538      0.742344   0.002272
2          0.249867      0.965363   0.002237
```

Table 7

Understanding Customer Segmentation:

Each cluster likely represents a distinct segment of the customer base. It will be great to analyse the characteristics of each cluster to understand the preferences, behaviours, and needs of the customers within each group.

Based on these clusters we can customise the email marketing fitting the customers' profile and offering them wine deals based on their preferences. It's important to consider the factors such as the types of wine, origin, campaign period, past peak and additional variables that form these clusters. This will help to create promotional offers that resonate most with each segment. The content and layout of the newsletter can also be adjusted to maximise engagement and sales for each segment.

To enhance our email marketing strategy, we can implement specific adjustments and closely monitor their impact on sales performance. This allows us to discern how different factors influence each cluster, enabling us to fine-tune our approach on various segments based on the specific needs and preferences of each customer segment.

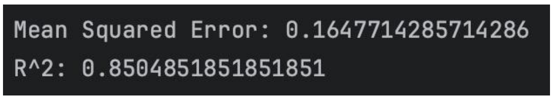
Model Building

Given the non-normality of the data and the identified non-linear relationships in "Minimum kg" and "Discount %," the model-building strategy involves employing Random Forests as a non-linear predictive technique for the dependent variable. This choice is substantiated by Random Forests' inherent ability to provide variable importance metrics, facilitating a nuanced understanding of each variable's contribution to the overall prediction.

Furthermore, Random Forests exhibit superior handling of categorical variables and demonstrate robustness to assumptions. While decision trees share similar qualities, the chance of overfitting is notably higher. Consequently, opting for Random Forests is the most optimal approach for this dataset.

For the random forest analysis, categorical variables in the dataset were encoded to facilitate the modeling process. Subsequently, the dataset was partitioned into training and testing sets, and the resulting model provided insights into the predictive relationships.

Additionally, the inclusion of Wine Clusters in the Random Forest model aims to enhance predictive accuracy by capturing complex relationships within the data. These clusters represent nuanced patterns not easily captured by individual features. If Wine Clusters exhibit predictive power and correlation with the target variable, integrating them enables the model to excel in capturing non-linear relationships, improving overall accuracy.



```
Mean Squared Error: 0.1647714285714286
R^2: 0.8504851851851851
```

Table 8

The obtained metrics indicate the following results, a low **MSE** of **0.16477** which indicates low average squared difference between the actual values and the values predicted by a model, and the high **R²** of **0.85** suggests that a significant portion of the variability in the target variable is accounted for by the model.

It is important to note that given the restricted dataset size, the model remains susceptible to overfitting. To advance research and facilitate a more robust identification of influential features, acquiring additional observations is crucial. Increased data volume would contribute to a more reliable and generalised model, enhancing its ability to discern and predict significant factors that affect wine preferences with greater accuracy.

For the model validation with only 32 observations, one effective technique is **k-fold cross-validation**. Given the limitations of a small sample size, splitting the data into multiple folds and iteratively training and testing the model on different subsets helps in obtaining more reliable performance estimates. This helps mitigate the risk of overfitting or underfitting that may occur with a small dataset.

Recommendations to enhance the sales:

Based on comprehensive research, various additional elements significantly shape wine preferences. Those are individual taste sensitivity, grape varietals, sweetness levels, food pairings, cultural influences, etc. Recognizing and addressing this spectrum of factors is important for the salesperson seeking to understand his consumer's preferences and enhance the amount of sales.

The salesperson should capitalise on the preference for aging in Sauvignon wines, strategically diversify the wine selection to include more aged options, and expand offerings in popular varietals like Chardonnay and Cabernet Sauvignon. Additionally, it's crucial to adjust import strategies to align with regional preferences, emphasising increased imports from favoured regions such as New Zealand, Chile and Oregon instead of France from where the exported wines cause relatively low sales. Following all these recommendations, the salesperson is likely to enhance his sales.

References:

1. Stanco, M., Lerro, M., & Marotta, G. (2020). Consumers’ Preferences for Wine Attributes: A Best-Worst Scaling Analysis. *Sustainability*, *12*(7), 2819. https://doi.org/10.3390/su12072819

2. Mehta, R., & Bhanja, N. (2017). Consumer preferences for wine attributes in an emerging market. *International Journal of Retail & Distribution Management*, *46*(1), 34–48. https://doi.org/10.1108/ijrdm-04-2017-0073

3. Borel, B. (2018, January 30). *Do your genes predict your wine preference?* Wine Enthusiast. https://www.wineenthusiast.com/culture/wine/supertasters-wine-preference/