PART 1:

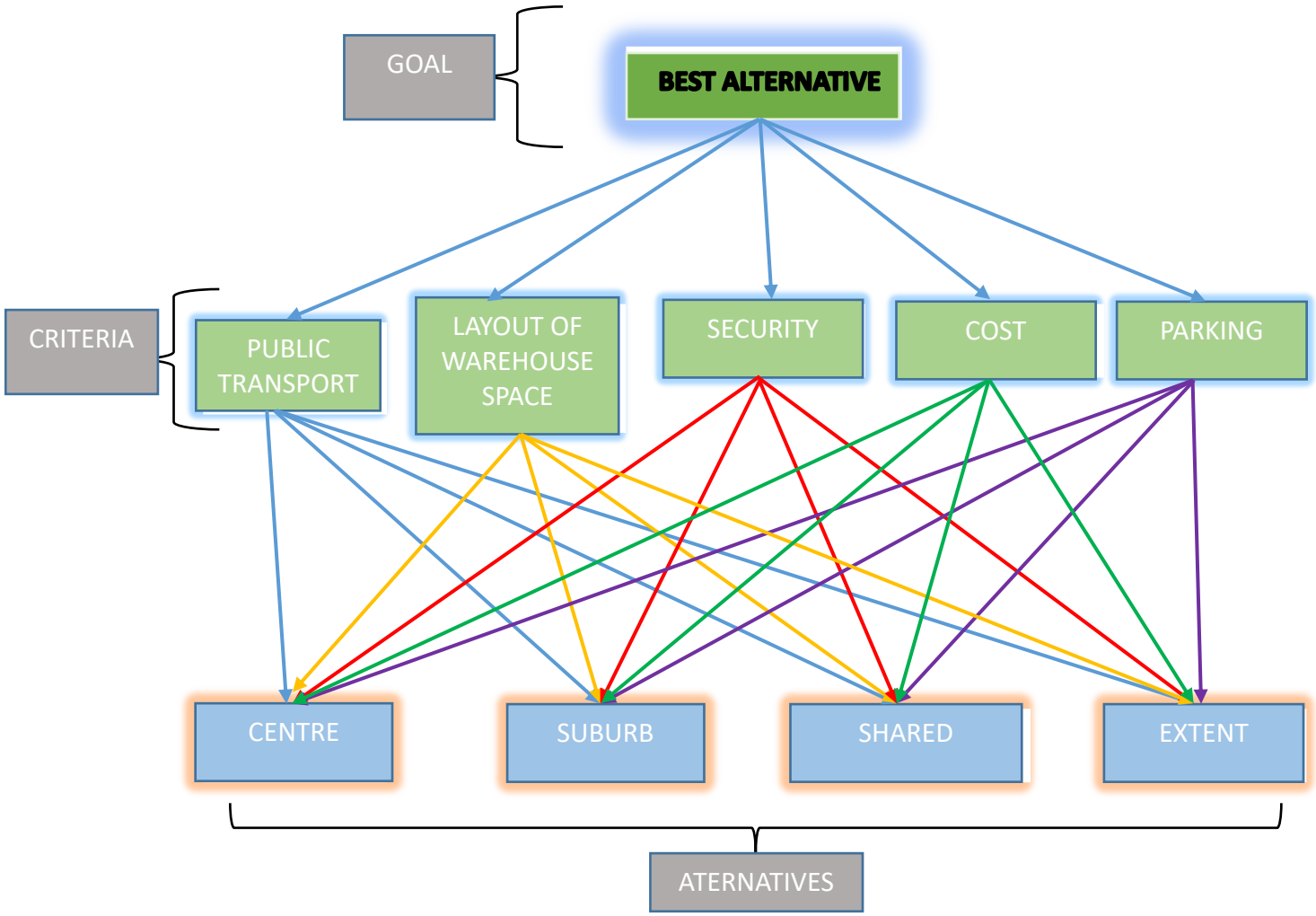
INTRODUCTION:

The objective of this report is to present, discuss, and apply the principles and techniques of the Analytic Hierarchy Process (AHP) and Technique for Order Preference by Similarity to Ideal Solutions (TOPSIS) in the prioritization and selection of the best way to acquire more space in a manufacturing company during expansion. The manufacturing company has 4 alternatives (Centre, Suburb, Shared and Extend), 5 criteria (Public transport links, Parking, Layout of the Warehouse space, Security and Cost). Table 1.1 shows about the information from the customer. The weights of the criteria used in the TOPSIS method are calculated using AHP.

Customer Input					
	C1 (public transport links)	**C2 (parking)**	**C3 (warehouse space)**	**C4 (security)**	**C5 (cost)**
A1 (Centre)	Good bus and rail links	Poor	Poor	***	£900,000
A2 (Suburb)	Good bus links but no rail links	Good	Excellent	****	£600,000
A3 (Shared)	Poor bus links but good rail links	Excellent	Good	***	£300,000
A4 (Extend)	Excellent bus and rail links	Moderate	Good	*	£200,000

Table 1.1: Information from the customer

BUSINUSS MODEL



SUMMARY:

Analytic Hierarchy Process:

AHP is one of the most prominent pairwise comparison methods. It uses a hierarchical structure of criteria and alternatives to organize a structure for MCDA problems. The following three steps includes the process of finding the best option among four alternatives using AHP.

1. **Derive the weightage:** Deriving the priorities is about allocating weightage to the criteria. The intensity of importance can be determined by using the Saaty scale.

   The pairwise comparison matrix (Table 1.2) consists of comparison data. When we weigh between parking and transport link, the value is 3 as I am giving moderate importance to parking than public transport (parking vs public transport). Moreover, if we compare public transport vs parking, the value is 1/3. Similarly the entire table is filled with specific weights and reciprocal of those values. All the grey cells represent the reciprocal values and, in addition to that, the diagonal values are always 1, since there is an equal importance.

Pairwise Comparision Matrix					
	C1 (public transport links)	C2 (parking)	C3 (warehouse space)	C4 (security)	C5 (cost)
C1 (public transport links)	1.00	0.33	0.25	0.20	0.17
C2 (parking)	3.00	1.00	0.50	0.25	0.20
C3 (warehouse space)	4.00	2.00	1.00	0.50	0.25
C4 (security)	5.00	4.00	2.00	1.00	0.50
C5 (cost)	6.00	5.00	4.00	2.00	1.00

Table 1.2: Pairwise Comparison Matrix

   The weightage table is driven after normalizing the pairwise matrix. The normalisation method and matrix multiplication are available in the Excel calculation work file. Table 1.3 shows the weights.

Weights	
Public transport links	*4.6%*
Parking	*8.9%*
Warehouse space	*14.9%*
Security	*27.1%*
Cost	*44.6%*

Table 1.3:

2. **Consistency check:** The next step is to look for any data inconsistencies. It means to checking whether the assigned weights are correct or not. This can be calculated using a consistency ratio (CR), the formula for CR = CI / RI.

CI Formula	$Consistency\ Index = \frac{\lambda_{max} - n}{n - 1}$

Table: Random Index

Attributes	3	4	5	6	7	8	9	10
RI	0.52	0.89	1.11	1.25	1.35	1.4	1.45	1.49

Table 1.4: Random Index

In the CI formula, n denotes the number of evaluated criteria (5) and lambda max is the maximum Eigenvalue. The max Eigenvalue is 5.25 from a normalised matrix table and RI is 1.11 from Random Index Table 1.4 for five criteria.
The consistency ratio (CR=CI/RI) is 0.03. Since its value is less than 10%, the considered data is more consistent.

3. **Making a final decision:** The given customer table has one quantitative data column (cost) and all the remaining are qualitative data types. In order make analysis easier, I will convert the qualitative data column by providing a rating from 1 to 4, 1 is for poor and 4 is excellent. As per the weightage that I got from the AHP process, I should choose the alternative which has minimum or average cost value and good transport, parking, warehouse space and security value.

Weightage	4.6%	8.9%	14.9%	27.1%	44.6%
	C1 (public transport links)	**C2 (parking)**	**C3 (warehouse space)**	**C4 (security)**	**C5 (cost)**
A1 (Centre)	3	1	1	3	9,00,000
A2 (Suburb)	2	3	4	4	6,00,000
A3 (Shared)	2	4	3	3	3,00,000
A4 (Extend)	4	2	3	1	2,00,000

Table 1.5: Input Data

As we can see in the data table (Table 1.5), each alternative is at its best at one or two criteria and if we try to make a decision as per that it's not easy to solve problem, so it is better to create a normalized matrix table to make a possible decision. Furthermore, take a lower value from the cost column and divide it by all the remaining values for the all alternatives corresponding row value. Next, identify the maximum value and divide the maximum value into remaining values. Finally, I will get the synthesis model by adding a multiplication of normalized values and criteria weightage for each alternative.
The synthesis model (Table 1.6) has AHP values to decide which alternative is the best. The alternative A3 (Shared) has highest value and it is considered as the best option among the four and the graph shows the level of AHP values for all the options.

Synthesis of model	
A1 (Centre)	0.396
A2 (Suburb)	0.658
A3 (Shared)	0.724
A4 (Extend)	0.715

Table 1.6:

TOPSIS:

TOPSIS is a multi-criteria decision analysis method based on the concept that the chosen alternative should have the shortest geometric distance from the positive ideal solution and the longest geometric distance from the negative ideal solution.

Our first step is to normalise the decision matrix using vector normalisation using the below formula, where xij is a score of alternative i with respect to criterion j. Table 1.7 has a vector normalized value.

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_i x_{ij}^2}}$$

	C1 (public transport links)	C2 (parking)	C3 (warehouse space)	C4 (security)	C5 (cost)
A1 (Centre)	0.522	0.183	0.169	0.507	0.789
A2 (Suburb)	0.348	0.548	0.676	0.676	0.526
A3 (Shared)	0.348	0.730	0.507	0.507	0.263
A4 (Extend)	0.696	0.365	0.507	0.169	0.175

Table 1.7: Vector Normalized Matrix

Now I will multiply the normalised values by the criteria's weights to get a weighted normalised matrix. Table 1.8 has a weighted normalised matrix for all the alternatives.

Weightage	4.6%	8.9%	14.9%	27.1%	44.6%
	C1 (public transport links)	C2 (parking)	C3 (warehouse space)	C4 (security)	C5 (cost)
A1 (Centre)	0.024	0.016	0.025	0.137	0.352
A2 (Suburb)	0.016	0.049	0.101	0.183	0.234
A3 (Shared)	0.016	0.065	0.075	0.137	0.117
A4 (Extend)	0.032	0.032	0.075	0.046	0.078

Table 1.8: Weighted Normalized Decision Matrix

The next stage is to calculate the ideal best and ideal worst solution. For non-beneficial criteria like cost, the ideal best solution is the lower value that is desired (0.078) and for all beneficial criteria the maximum value is desired (maximum values 0.183, 0.101, 0.065, 0.032 for security, warehouse space, parking, public transport links respectively) and vice versa for ideal worst solution.

The best alternative is the one that is closest to the positive ideal solution (PIS) and furthest away from the positive ideal solution (NIS); to find that, I calculate the Euclidian distance between each alternative and PIS / NIS.
To start, I calculate the distance between each alternative to the PIS using the following formula: where S* is the Euclidean distance from PIS and S- is the Euclidean distance from NIS. Table 1.9 shows PIN, NIS and Euclidean ideal best and worst solution.

$$S_i^* = \sqrt{\sum \left(V_{ij} - V_j^*\right)^2} \qquad S_i^- = \sqrt{\sum \left(V_{ij} - V_j^-\right)^2}$$

Weightage	4.6%	8.9%	14.9%	27.1%	44.6%	Euclidean Distance	
	C1 (public transport links)	C2 (parking)	C3 (warehouse space)	C4 (security)	C5 (cost)	Ideal Best (S*)	Ideal Worst (S-)
A1 (Centre)	0.023957019	0.016191259	0.025164886	0.137442874	0.351675471	0.291612246	0.091975911
A2 (Suburb)	0.015971346	0.048573777	0.100659546	0.183257165	0.234450314	0.157946181	0.198444834
A3 (Shared)	0.015971346	0.064765036	0.075494659	0.137442874	0.117225157	0.067187532	0.261257047
A4 (Extend)	0.031942692	0.032382518	0.075494659	0.045814291	0.078150105	0.143430968	0.279045611
Ideal Best Value (V*)	0.031942692	0.064765036	0.100659546	0.183257165	0.078150105		
Ideal Worst value (V-)	0.015971346	0.016191259	0.025164886	0.045814291	0.351675471		

Table 1.9:

Finally, I can calculate the relative closeness to the ideal solution using the following calculation, where the Final score (P) is a TOPSIS value which denotes the best alternative option.

$$Final\ Score = S_i^- / \left(S_i^* + S_i^-\right)$$

Alternatives	Final Score (P)
A1 (Centre)	0.24
A2 (Suburb)	0.56
A3 (Shared)	0.80
A4 (Extend)	0.66

Table 2.1:

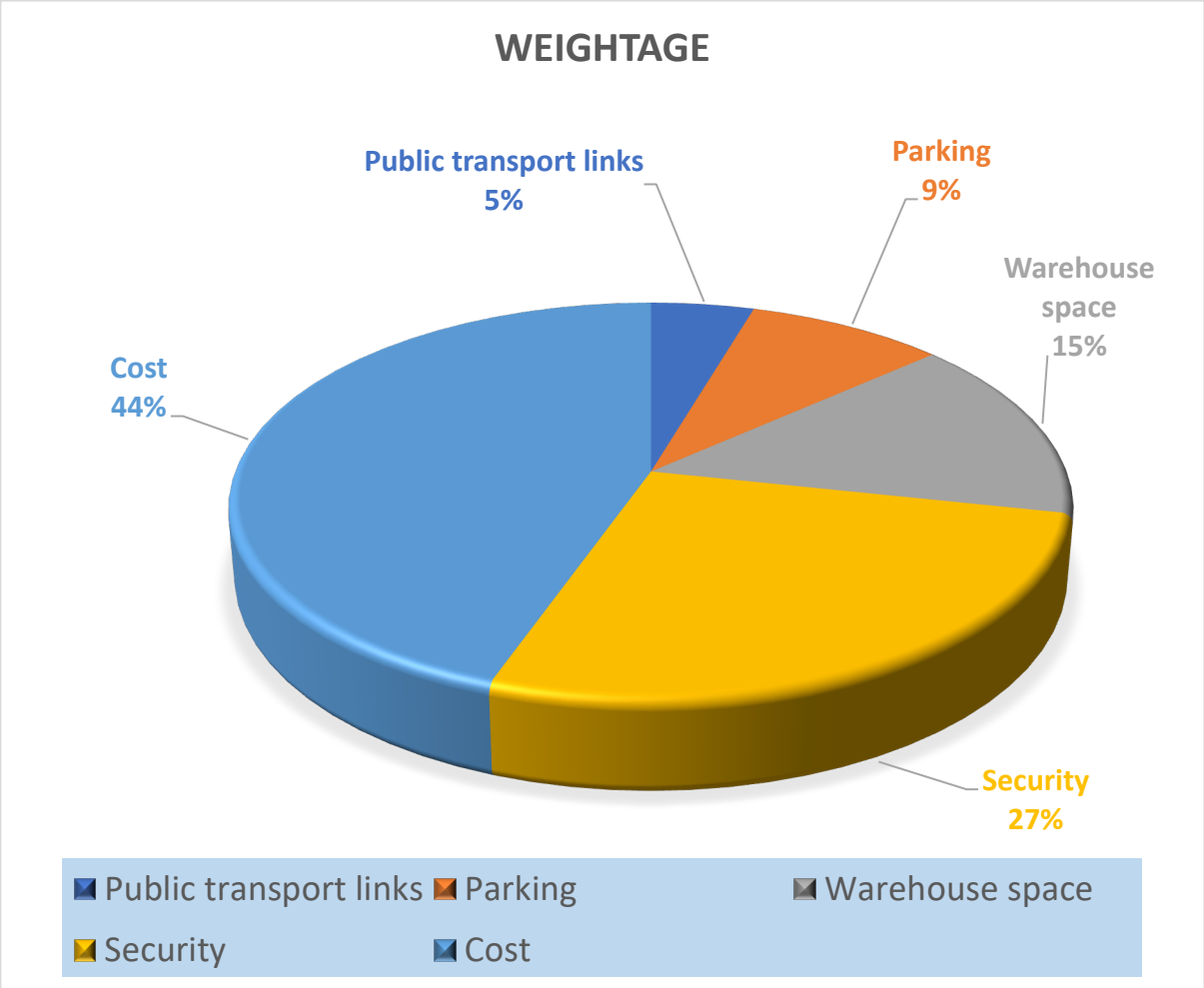
Table 2.1 has the final score of alternatives. As per TOPSIS, the higher the score, the better the alternative, the Shared (A3) option is the best performing alternative!

RESPONSE TO COMPANY:

Overall, the shared alternative (A3) is the best option from both methods. The Shared alternative has 3* security, excellent parking, good transport facility, good layout of warehouse space and average cost required.

The weights calculated from the AHP process clearly show that I am giving highest priority to cost (44%), followed by security (27%), layout of warehouse space (15%), parking (9%) and public transport links (5%). However, the percentage is, in its totality, a cognitive and mental process derived from the most possible adequate selection based on tangible and intangible criteria, which are arbitrarily chosen by those who make the decisions.

The cost is the main criterion here that I am considering to ensure that the expenditure amount should not exceed the allocated budget. Secondly, I am giving next preference to security because risk is unexpected, so it is always better to be on the safer side.



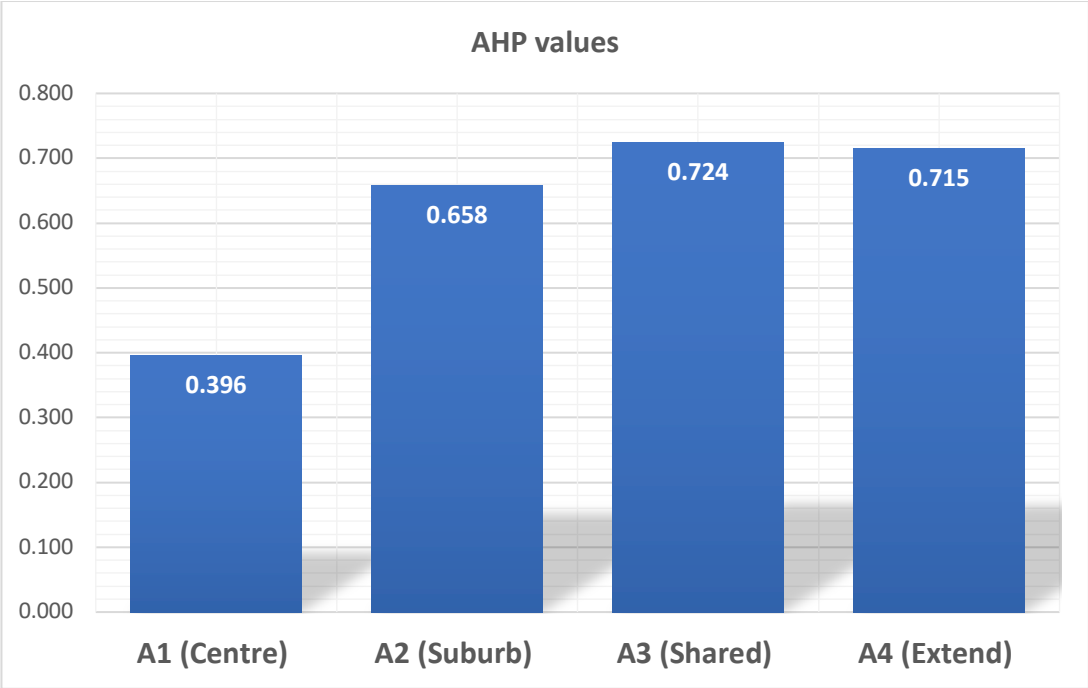
Graph 1.1: Pie Chart

The consistency ratio (CI) is a metric that indicates the consistency between pairwise comparisons. It is a measurement that indicates how much you deviate from the

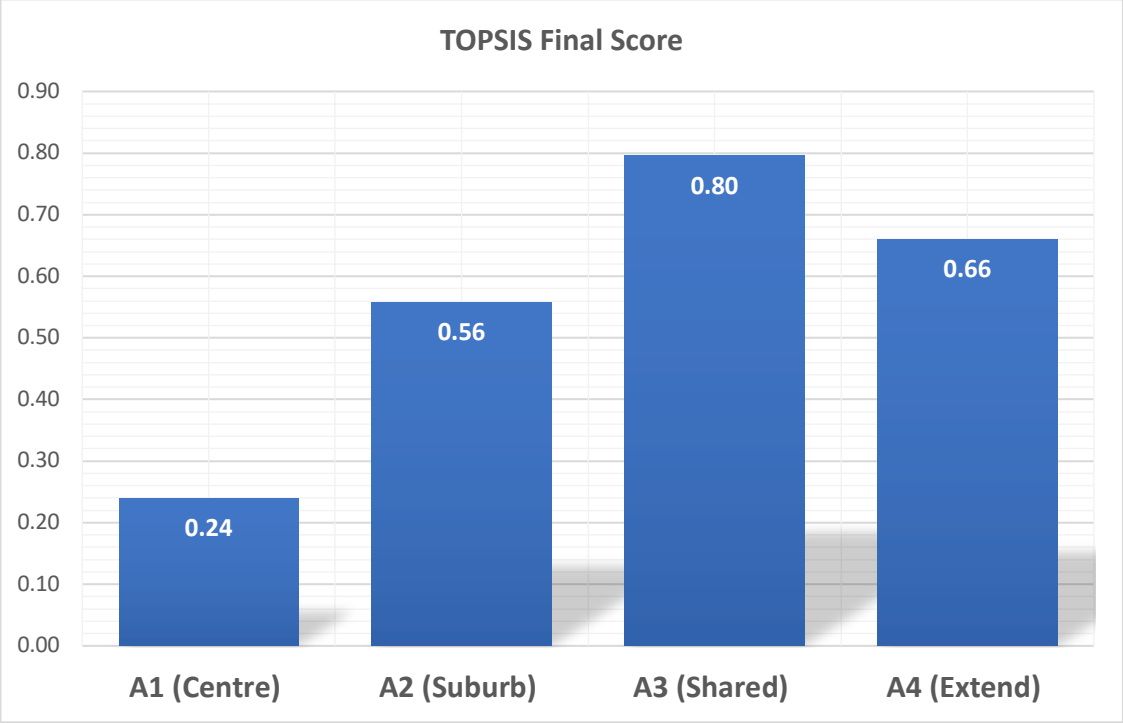
consistency. According to Saaty, if it is less than 10% the pairwise comparisons are accepted and the weightage can be used for the analysis part.

The pie chart "WEIGHTAGE" shows the percentage distributed on each criterion. The sum of all the percentages must be 1. The weightage is used for the AHP and the TOPSIS method is the same as it obeying the consistency ration check.

Using AHP, I can clearly recommend that A3 is the first preferred option as its value is in the top position (0.724) and A4 is the next immediate choice (0.715) but in the case of TOSIS, there is only alternative solution which is A3 as its value 0.8 and there is considerable difference between other alternatives values. Graph 1.2 and Graph 1.3 show alternative AHP and TOPSIS values with respect to each other.



Graph 1.2: AHP Values



Graph 1.3: TOPSIS Score

PART 2:

INTRODUCTION:

The purpose of this report is to analyse a small dataset for Brewdog Beer company. The report includes dealing with missing data and imputation to refine the data to create subsets of data with similar attributes. The missing data is perhaps one of the most common issues when dealing with any dataset and there are many reasons why data could be missing. However, handling missing data plays a major role in data analysis, since if the data quality is not up to the mark, it leads to distorted results. In this report I will explain various types of missing data, methods to handle missing data and the clustering in a hierarchical algorithm with dendrogram.

MISSING DATA AND HANDLING MISSING DATA:

In the given dataset there are 199 observations and 9 variables. Seven variables contain quantitative data, namely Alcohol by volume (ABV), International Bitterness Units (IBU), Original Gravity (OG), Colour Units from the European Brewery Convention (EBC), pH (Acid & Base Scale), Attenuation Level and Fermentation temperature in Celsius.

The percentage of missing data is 5.5%. Overall 11 observations have missing data. To be clear, 7 elements from ABV and 4 from EBC.

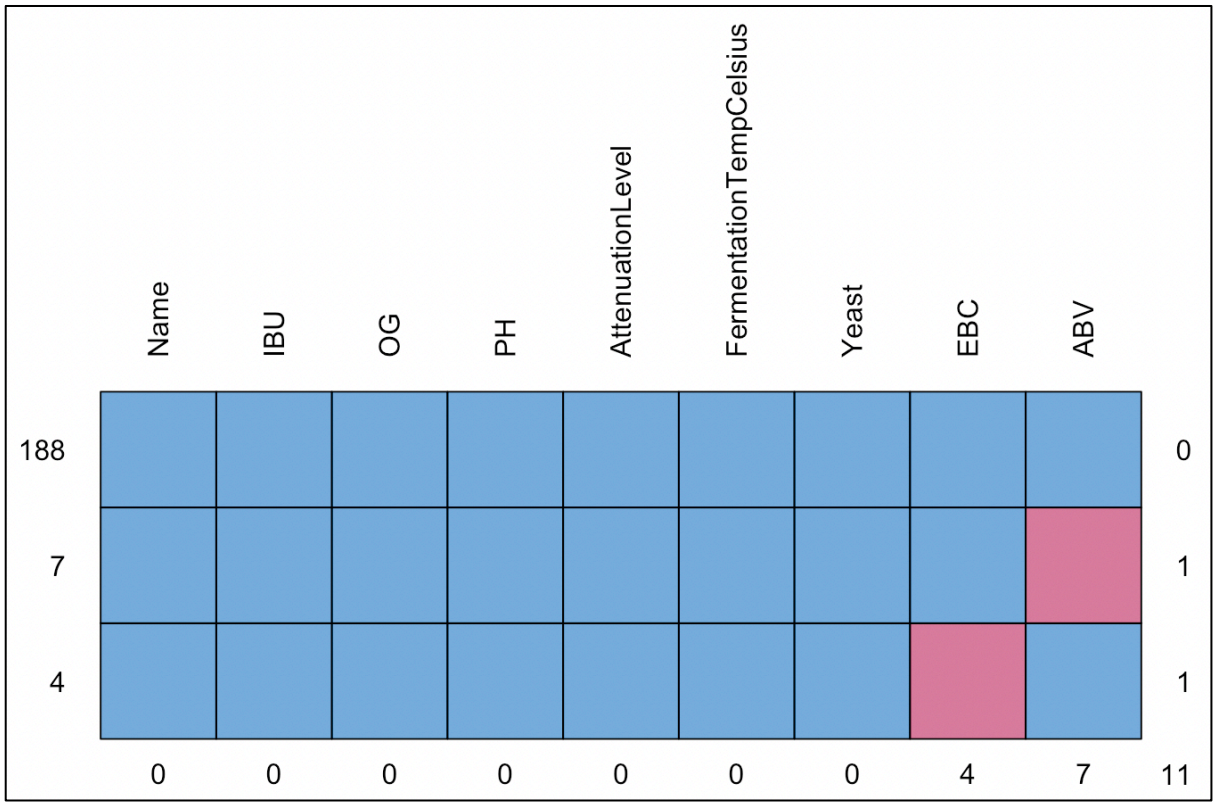


Figure 1.1: Missing Data Pattern

Figure 1.1 shows the pattern of missing data; the blue coloured cells indicate correct data and the light red coloured cells specify missing data. Figure 1.2 demonstrates the combination of missing data.

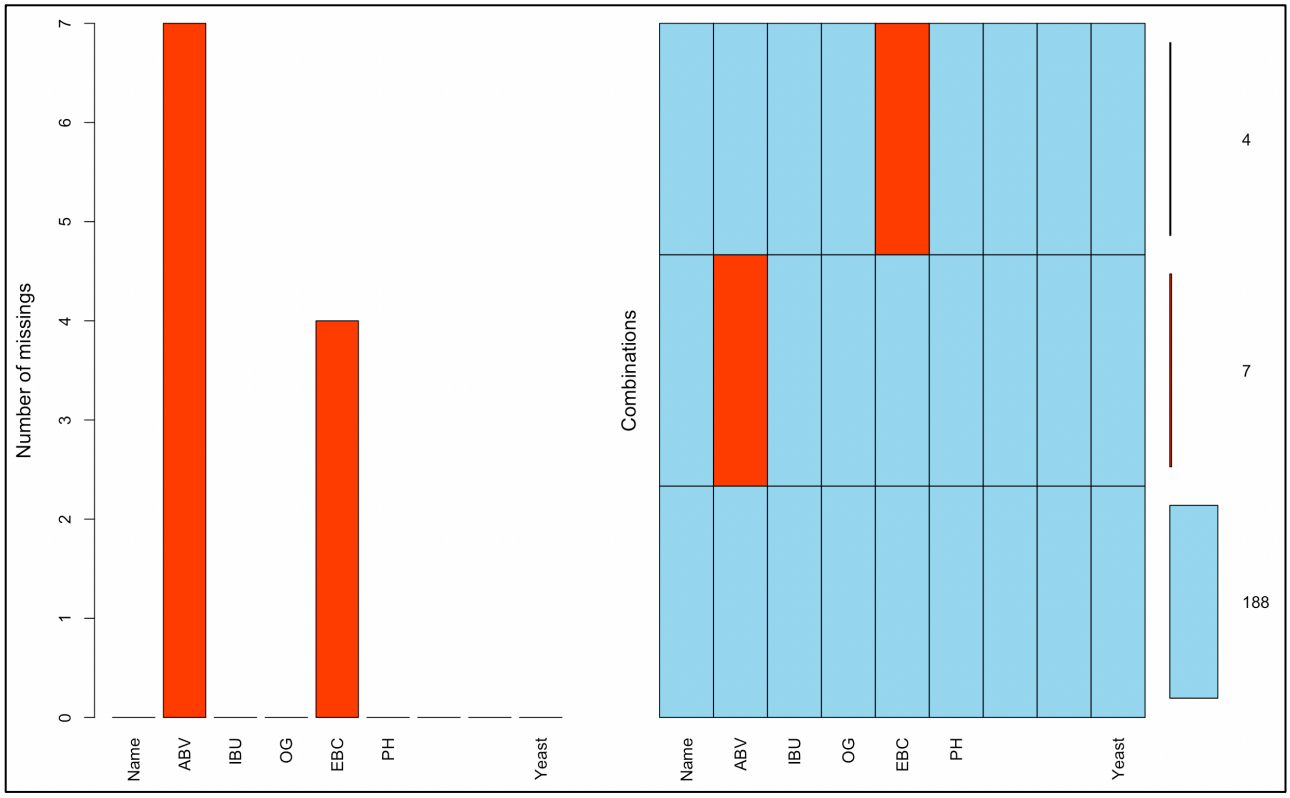


Figure 1.2: Combination Graph

The combination graph 1.2 on the right shows the red blocks which highlight where information is missing and the scale on the right shows the number of missing records for that combination. As we can see in the figure, no observation has more than one missing element.

Before deciding which approach to employ for data handling, we should understand why the data is missing. There will be three possible ways,

- Missing at Random (MAR): which means the data is missing relative to the observed data.
- Missing Completely at Random (MCAR): In the MCAR situation, the data is missing across all observations regardless of the expected value or other variables.
- Missing Not at Random (MNAR): The MNAR category applies when the missing data has a structure to it. In other words, there appear to be reasons the data is missing.

In the current beer dataset, the missing data completely belongs to MCAR category. When dealing with missing data, we can use two primary methods to solve the error. One is imputation and the other deletion. The imputation method develops reasonable guesses for missing data. The other option is to remove data or deletion. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. Deciding

between deletion and imputation depends on which outcome we think will produce the most reliable and accurate results.

1. **Deletion:**

There are three primary methods for deleting data when dealing with missing data: listwise, pairwise and dropping variables.

- **Listwise and Pairwise:**

  In this method, all data for an observation that has one or more missing values is deleted. The analysis is run only on observations that have a complete set of data. If the data set is small, it may be the most efficient method to eliminate those cases from the analysis. However, in most cases, the data is not missing completely at random (MCAR).

- **Dropping Variables:**

  If data is missing for more than 60% of the observations, it may be wise to discard it if the variable is insignificant.

I am not going with deletion approach as the data missing percentage is much smaller (6%).

2. **Imputation:**

Instead of deletion, we have multiple solutions to improve the value of missing data. Depending on why the data is missing, imputation methods can deliver reasonably reliable results. These are two types of imputation.

- **Single Imputation (Mean, Median and Mode):**

  This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations, we can calculate the mean or median of the existing observations.
  This method is more suitable in the current scenario. If we neglect NA in the ABV column, the mean is 7.675. This value will be replace by the NA value to get complete data cases. After replacing the mean by NA values, the overall mean value of the variable ABV does not changes much and still with the same value at 7.675. Similarly, after replacing the mean value by NA of the variable EBC, the mean is 71.66. Now, the dataset is complete.

- **Multiple Imputations:**

  Multiple imputations are considered a good approach for data sets with a large amount of missing data. Instead of substituting a single value for each missing data point, the missing values are exchanged for values that encompass the natural variability and uncertainty of the right values. Using the imputed data, the process is repeated to make multiple imputed data sets. Each set is then analysed using the standard analytical procedures, and the multiple analysis results are combined to

produce an overall result. The various imputations incorporate natural variability into the missing values, which creates a valid statistical inference. Multiple imputations can produce statistically valid results even when there is a small sample size or a large amount of missing data.

CLUSTERING ALGORITHM AND DENDROGRAM:

Clustering is basically a technique that groups similar data points such that the points in the same group are more similar to each other than the points in the other groups. The group of similar data points is called a cluster. In the given dataset, we have 199 types of beer, we need to create a cluster based on the dependent variables ABV, IBU, OG, EBC, pH, Attenuation Level, Fermentation temperature and Yeast type. The clustering can be done easily using packages fastcluster, NbClust and cluster in R.

There are mainly two types on clustering, the non-hierarchical clustering and hierarchical clustering.

- **Non-hierarchical clustering:**

  Cluster analysis with a non-hierarchical method is a clustering method that manually determines the number of clusters. The non-hierarchical cluster analysis technique is designed to group items, not variables, into clusters of K clusters. The number of clusters, K, is predefined to start the clustering procedure. The non-hierarchical cluster analysis method is related to K-Means. K-Means clustering is very suitable for large data sizes because it has a higher speed than the hierarchical method. It is a distance-based clustering method that divides data into a number of clusters and only works on numeric attributes.

- **Hierarchical clustering:**

  Hierarchical Clustering Algorithm, also known as Hierarchical cluster analysis or HCA, is an unsupervised clustering algorithm which involves creating clusters that have predominant ordering from top to bottom. The endpoint is a set of clusters, where each cluster is distinct from each other, and the objects within each cluster are broadly similar to each other.

  The hierarchy relationships between all the clusters are represented using dendrogram. The dendrogram is a tree diagram used to visualize and classify taxonomic relationships frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering and it can represent any type of grouped data.

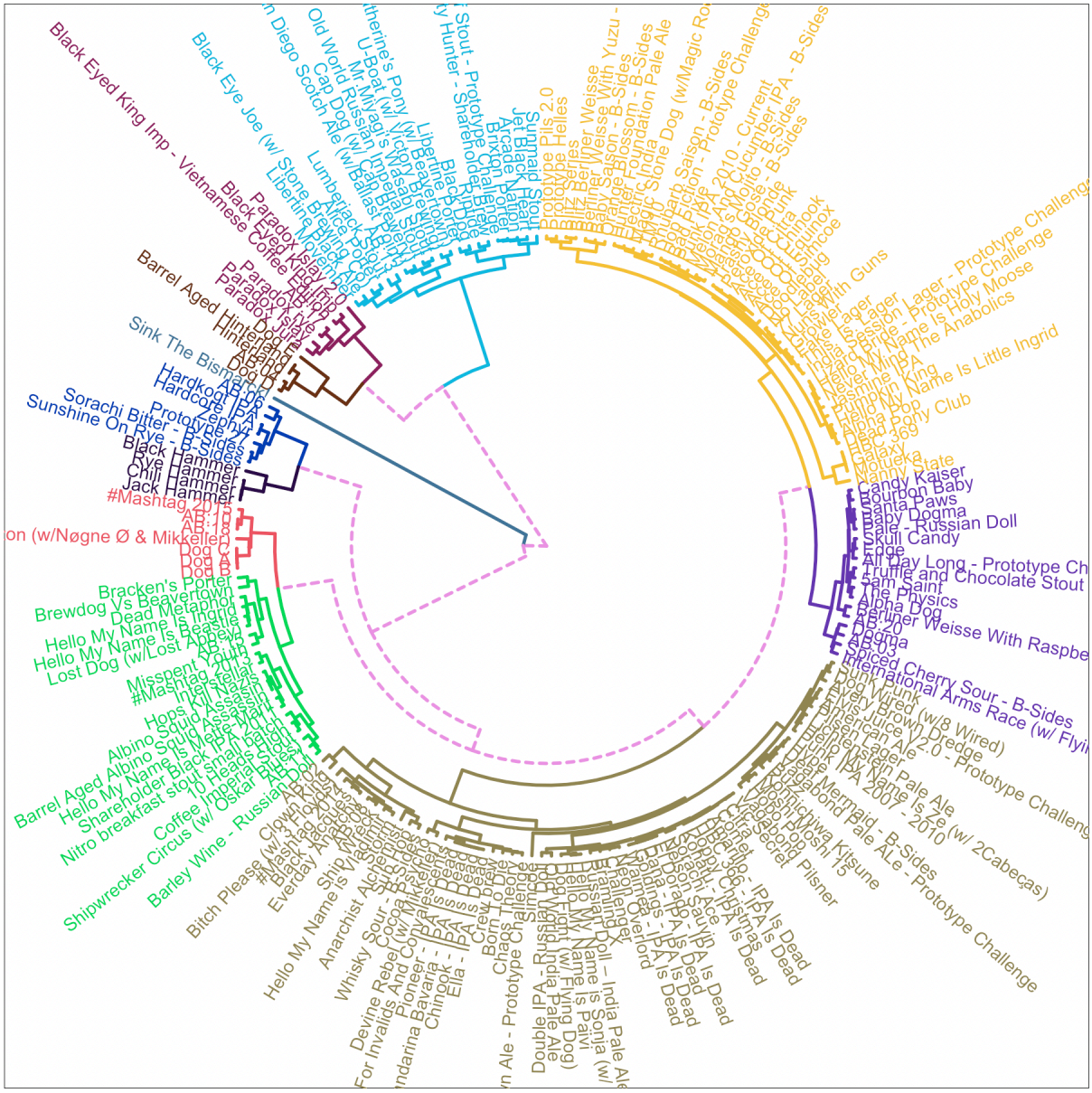


Figure 1.4: Customized Dendrogram (Fan type)

The dendrogram consists of stacked branches (called clades) that break down into further smaller branches. At the lowest level, there will be individual elements and then they are grouped according to attributes into clusters with fewer and fewer clusters at higher levels. The arrangement of the beers reveals how similar they are to each other; two beers in the same clade are more similar than two beers in another clade. The y-axis (the height of the branch) shows how close data points or clusters are to one another. The taller the branch, the further and more different the clusters. Table 1.1 shows the clustered beers which are segregated as per dependent variable values.

Clusters	Beers
1	#Mashtag 2013, 10 Heads High, AB:11, AB:15, Albino Squid Assasin, Barley Wine - Russian Doll, Barrel Aged Albino Squid Assassin, Bracken's Porter, Brewdog Vs Beavertown, Coffee Imperial Stout, Dead Metaphor, Hello My Name Is Beastie, Hello My Name Is Ingrid, Hello My Name Is Mette-Marit, Hops Kill Nazis, Interstellar, Lost Dog (w/Lost Abbey), Misspent Youth, Nitro breakfast stout small batch, Shareholder Black IPA 2011, Shipwrecker Circus (w/ Oskar Blues)
2	#Mashtag 2014, AB:02, AB:08, Amarillo - IPA Is Dead, American Ale, Anarchist Alchemist, Avery Brown Dredge, Bitch Please (w/ 3 Floyds), Black Jacques, Born To Die, Bramling X, Buzz, Challenger, Chaos Theory, Chinook - IPA Is Dead, Citra, Clown King, Cocoa Psycho, Comet, Crew brew, Dana - IPA Is Dead, Deaf Mermaid - B-Sides, Devine Rebel (w/ Mikkeller), Dog Fight (w/ Flying Dog), Dog Wired (w/8 Wired), Double IPA - Russian Doll, El Dorado - IPA Is Dead, Ella - IPA Is Dead, Elvis Juice V2.0 - Prototype Challenge, Everday Anarchy, Goldings - IPA Is Dead, HBC 366 - IPA Is Dead, Hello My Name Is Päivi, Hello My Name is Sonja (w/ Evil Twin), Hello My Name is Vladimir, Hello My Name Is Zé (w/ 2Cabeças), Hobo Pop, Hopped-Up Brown Ale - Prototype Challenge, Hoppy Christmas, Hype, Kohatu - IPA Is Dead, Konnichiwa Kitsune, Lichtenstein Pale Ale, Mandarina Bavaria - IPA Is Dead, Moshi Moshi 15, Nelson Sauvin, Neon Overlord, Old World India Pale Ale, Pilsen Lager, Pioneer - IPA Is Dead, Punk IPA 2007 - 2010, Restorative Beverage For Invalids And Convalescents, Russian Doll – India Pale Ale, Ship Wreck, Simcoe, Sorachi Ace, Storm, Sub Hop, Sunk Punk, Vagabond Pale ALe - Prototype Challenge, Vagabond Pilsner, Vic Secret, Waimea - IPA Is Dead, Whisky Sour - B-Sides
3	#Mashtag 2015, AB:10, AB:18, Black Tokyo Horizon (w/Nøgne Ø & Mikkeller), Dog A, Dog B, Dog C
4	5am Saint, AB:03, AB:20, All Day Long - Prototype Challenge, Alpha Dog, Baby Dogma, Berliner Weisse With Raspberries And Rhubarb - B-Sides, Bourbon Baby, Candy Kaiser, Dogma, Edge, International Arms Race (w/ Flying Dog), Pale - Russian Doll, Santa Paws, Skull Candy, Spiced Cherry Sour - B-Sides, The Physics, Truffle and Chocolate Stout - B-Sides
5	77 Lager, Ace Of Chinook, Ace Of Citra, Ace Of Equinox, Ace Of Simcoe, Alpha Pop, Baby Saison - B-Sides, Bad Pixie, Berliner Weisse With Yuzu - B-Sides, Blitz Berliner Weisse, Blitz Series, Dead Pony Club, Doodlebug, Electric India, Fake Lager, Galaxy, Growler, HBC 369, Hello My Name Is Holy Moose, Hello My Name Is Little Ingrid, Hop Fiction - Prototype Challenge, Hunter Foundation Pale Ale, India Session Lager - Prototype Challenge, Jasmine IPA, Lizard Bride - Prototype Challenge, Magic Stone Dog (w/Magic Rock & Stone Brewing Co.), Mango Gose - B-Sides, Melon And Cucumber IPA - B-Sides, Morag's Mojito - B-Sides, Motueka, Nanny State, Never Mind The Anabolics, No Label, Nuns With Guns, Orange Blossom - B-Sides, Peroxide Punk, Prototype Helles, Prototype Pils 2.0, Pumpkin King, Punk IPA 2010 - Current, Rhubarb Saison - B-Sides, This. Is. Lager, TM10, Trashy Blonde
6	AB:04, Barrel Aged Hinterland, Dog D, Dog E, Hinterland
7	AB:06, Hardcore IPA, Hardkogt IPA, Prototype 27, Sorachi Bitter - B-Sides, Sunshine On Rye - B-Sides, Zephyr
8	AB:13, Alice Porter, Arcade Nation, Black Dog, Black Eye Joe (w/ Stone Brewing Co), Bounty Hunter - Shareholder Brew, Brixton Porter, Cap Dog (w/ Cap Brewery), Catherine's Pony (w/ Beavertown), Jet Black Heart, Libertine Black Ale, Libertine Porter, Lumberjack Stout, Movember, Mr.Miyagi's Wasabi Stout, Old World Russian Imperial Stout, Riptide, San Diego Scotch Ale (w/Ballast Point), Stereo Wolf Stout - Prototype Challenge, Sunmaid Stout, U-Boat (w/ Victory Brewing)
9	AB:17, Black Eyed King Imp, Black Eyed King Imp - Vietnamese Coffee Edition, Paradox Islay, Paradox Islay 2.0, Paradox Jura, Paradox rye
10	Black Hammer, Chili Hammer, Jack Hammer, Rye Hammer
11	Sink The Bismarck!

Table 1.1: Clustered Beers

As we can see in Figure 1.4, the total 199 beers are divided into 11 clusters and each cluster represented in a unique colour. The cluster which is identified in light grey colour has a max number of beers, followed by a yellow colour subset. The group coloured with green and sky blue shared an equal number of beers and the purple coloured cluster comes next with 2 fewer beers. The beer “Sink The Bismarck!” moved to a separate cluster due to high ABV, IBU and OG values. Apart from that all other coloured clusters share approximately 4 to 6 beers equally. However, using a hierarchical clustering dendrogram, I can show all the possible linkage between clusters but the biggest con of the dendrogram is **scalability**. When there is a large set of data, it is extremely hard to examine and computationally expensive. Figure 1.5 shows the main reason why dendrograms are far from amazing. In order to show the tree in a more understandable way, I created fan type customized-dendrogram.

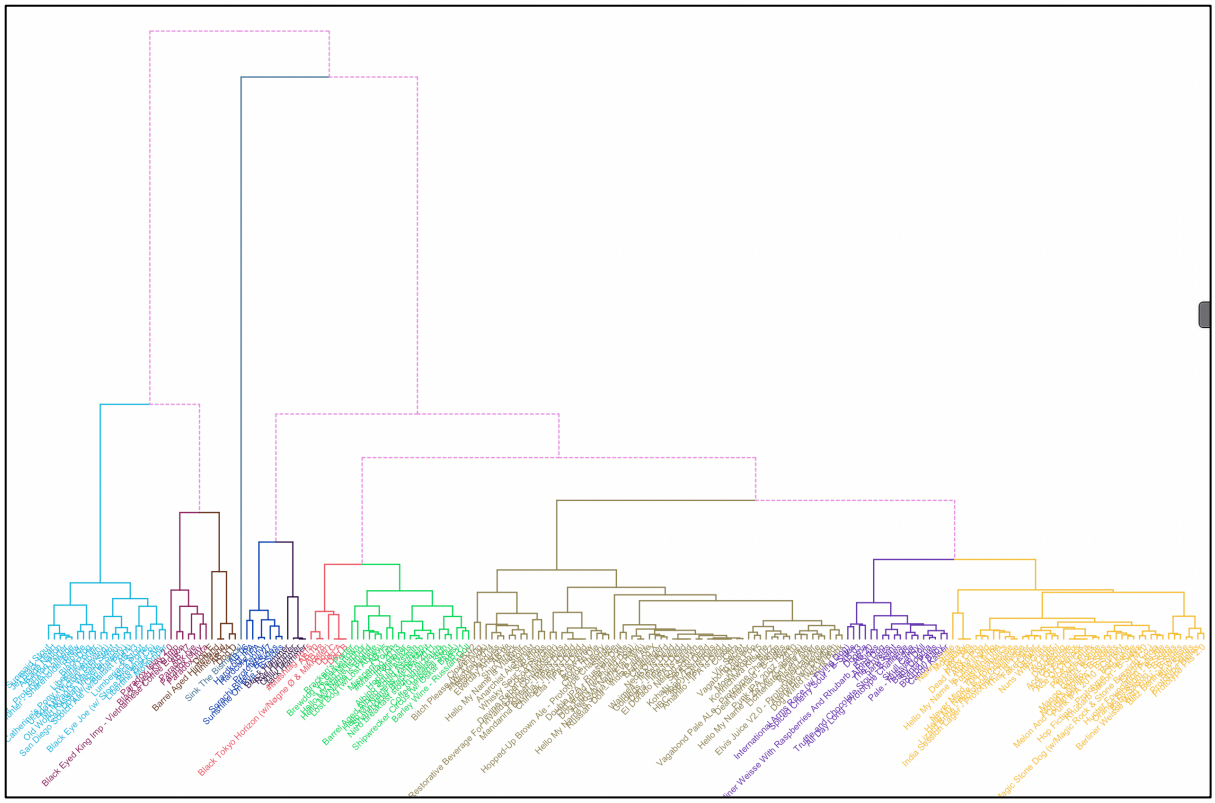


Figure 1.5: Dendrogram