Assessment Submission Cover Sheet

This Assessment Cover Sheet **must** be included on all Assessment submissions.

Assignment Title	Assignment B – Portfolio Assessment
Module	Data Mining
Student Name (same as Student Card)	Ciaran Finnegan
Student Number	Ciaran: D21124026
Programme	TU060
Part-Time/Full-Time	Part-time
Year of Study (First Year, Second Year, etc)	First Year

Late Submissions: Assessment submitted after the deadline will have a late penalty applied.

Academic Integrity for assessment in TU Dublin Programmes

Each student is responsible for knowing and abiding by TU Dublin Academic Regulations and Policies. Any student in breach of these regulation/policies will be subject to action in accordance with the University's procedures for breaches of assessment regulations. Please refer to the General Assessment Regulations at
https://tudublin.libguides.com/c.php?g=674049&p=4794713
https://www.tudublinsu.ie/advice/exams/breachesofregulations/

All students are expected to complete their courses/programmes in compliance with University regulations. No student shall engage in any activity that involves attempting to receive a grade by means other than honest effort, for example:

1. No student shall complete, in part or in total, any examination or assessment for another person.
2. No student shall knowingly allow any examination or assessment to be completed, in part or in total, for themselves by another person.
3. No student shall plagiarise or copy the work of another and submit it as their own work.
4. No student shall falsify any data. Falsification is the invention of data, its alteration, its copying from any other source, or otherwise obtaining it by unfair means, or inventing quotations and/or references.
5. No student shall use aids or devices excluded by the lecturer in undertaking course work or assessments/ examinations.
6. No student shall knowingly procure, provide, or accept any materials that contain questions or answers to any examination or assessment to be given at a subsequent time.
7. No student shall provide their assignments, in part or in total, to any other student in current or future classes of this module/ programme unless authorised to do so by the lecturer.
8. No student shall submit substantially the same material in more than one module/programme without prior authorization.
9. No student shall alter graded assignments or examinations and then resubmit them for regrading, unless specifically authorised to do so by the lecturer.
10. All programming code and documentation, unless correctly referenced, submitted for assessment, or existing in the student's computer accounts must be the students' original work or material specifically authorized by the lecturer.
11. Collaborating with other students to develop, complete or correct course work is limited to activities explicitly authorized by the lecturer.
12. For all group assignments, each member of the group is responsible for the academic integrity of the entire submission. Consequently, all group members must satisfy themselves that all elements of their submission adhere to the academic integrity statement points above.

By submitting coursework, either physically or electronically, you are confirming that it is your own work (or, in the case of a group submission, that it is the result of joint work undertaken by members of the group that you represent) and that you have read and understand the University's Regulations and Policies covering Academic Integrity (see General Assessment Regulations).

Coursework may be submitted to an electronic detection system in order to help ascertain if any plagiarised material is present. If you have queries about what constitutes plagiarism, please speak to your lecturer.

Student Signature	Ciaran Finnegan
Date	8th January 2022

Submission of an Assessment, either physically or electronically, with or without this cover sheet, acknowledges your compliance with the TU Dublin Academic Regulations and Policies.

IMPORTANT:

- Complete the required number of tasks as defined in Assessment Handout
- The sections listed below are an example of the section headings for each task. You can use alternative headings
- Tasks 1-3: Sub-Sections 1-7 should be no longer than 8 pages (minimum 6 pages), including diagrams, images, screen captures, tables, etc. Careful selection of these is needed.
    - Code does not count to this total. Code should be added to the relevant section.
- Detailed discussion is expected. Marks are awarded based on depth of information given.
- Marks are awarded based on complexity of problem and depth of work.

TASK 1 – *Clustering: Analysis of Craft Beer Recipe Dataset to isolate preferred IPA recipes and brewing process.*

1. Definition of Problem

The objective of this task is to look at publicly available homebrew recipes for craft beer and determine if patterns can be established to isolate the American IPA beer recipes most likely to favour the following characteristics:

- Stronger than average alcohol by volume (ABV).
- Generally, more bitter in taste (scores higher on the 'International Bittering Units' – IBU – scale).
- Darker colour (just a personal preference).

A website called the Brewer's Friend allows homebrew enthusiasts to upload and share their own recipes. A Kaggle project is located here: Brewer's Friend Beer Recipes | Kaggle, which has scraped most of the recipe information into a dataset of 75,000 records of homebrew beers.

The investigation/output criteria listed in the bullet points above reflect my personal preference. The ideal outcome for this assignment task is to assess if clusters/segments exist in the recipe dataset that represent a brewing process, which I can try out domestically, that is most likely to deliver my desired type of American IPA homebrew beer.

To conduct this analysis, I downloaded the 14Mb homebrew recipe dataset from Kaggle and ran a parallel set of clustering investigations using both SAS Enterprise Miner and a small Python program, written in Jupyter Notebooks.

This complimentary approach allowed me to take advantage of the visual and data outputs from the '*Cluster*' and '*Segment Profile*' nodes in SAS EM, while also having a logical basis for the numbers of clusters chosen – based on the Python code that ran a KMeans analysis on the filtered dataset.

In this task report I will alternate between SAS EM and Python screenshots, depending on which format is best suited to represent information.

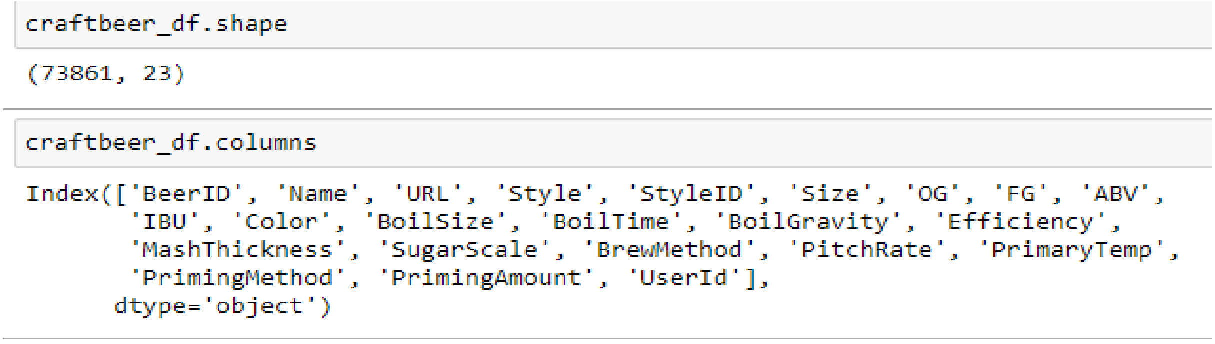
2. Data Exploration & Descriptive Analytics

Basic Dataset Dimensions and Quality

The dimension of the craft beer homebrew recipe dataset is:

- 73, 861 rows
- 23 columns
- 12 numerical features
- 11 categorical columns

A quick Python generated snapshot of the dataset shows the following columns:



```
craftbeer_df.shape
```

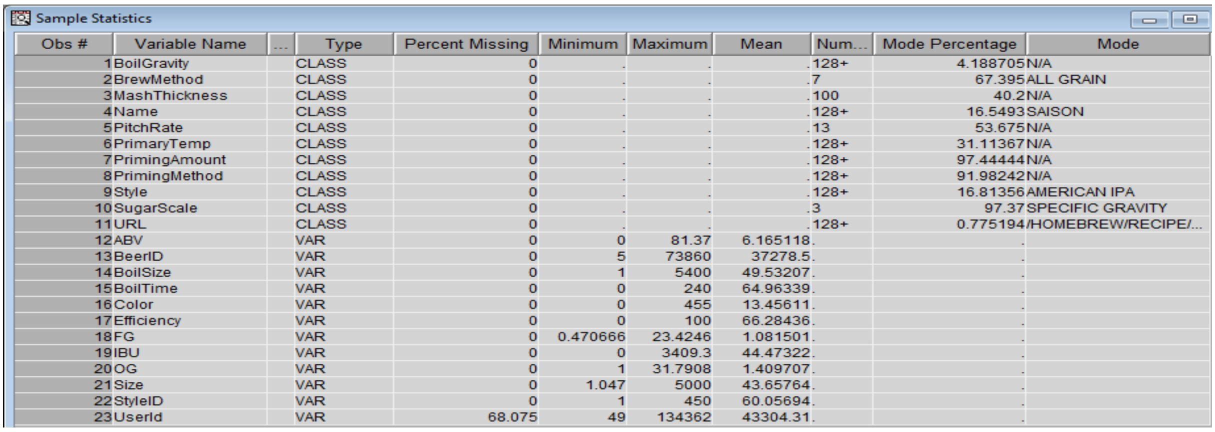
```
(73861, 23)
```

```
craftbeer_df.columns
```

```
Index(['BeerID', 'Name', 'URL', 'Style', 'StyleID', 'Size', 'OG', 'FG', 'ABV',
       'IBU', 'Color', 'BoilSize', 'BoilTime', 'BoilGravity', 'Efficiency',
       'MashThickness', 'SugarScale', 'BrewMethod', 'PitchRate', 'PrimaryTemp',
       'PrimingMethod', 'PrimingAmount', 'UserId'],
      dtype='object')
```

Fig 1. Python – Dataset Dimensions and Colum List

Looking at the attributes in SAS EM provides more detail on data quality:



Sample Statistics

Obs #	Variable Name	...	Type	Percent Missing	Minimum	Maximum	Mean	Num...	Mode Percentage	Mode
1	BoilGravity		CLASS	0	.	.	.	128+	4.188705	N/A
2	BrewMethod		CLASS	0	.	.	.	7	67.395	ALL GRAIN
3	MashThickness		CLASS	0	.	.	.	100	40.2	N/A
4	Name		CLASS	0	.	.	.	128+	16.5493	SAISON
5	PitchRate		CLASS	0	.	.	.	13	53.675	N/A
6	PrimaryTemp		CLASS	0	.	.	.	128+	31.11367	N/A
7	PrimingAmount		CLASS	0	.	.	.	128+	97.44444	N/A
8	PrimingMethod		CLASS	0	.	.	.	128+	91.98242	N/A
9	Style		CLASS	0	.	.	.	128+	16.81356	AMERICAN IPA
10	SugarScale		CLASS	0	.	.	.	3	97.37	SPECIFIC GRAVITY
11	URL		CLASS	0	.	.	.	128+	0.775194	/HOMEBREW/RECIPE/...
12	ABV		VAR	0	0	81.37	6.165118	.	.	
13	BeerID		VAR	0	5	73860	37278.5	.	.	
14	BoilSize		VAR	0	1	5400	49.53207	.	.	
15	BoilTime		VAR	0	0	240	64.96339	.	.	
16	Color		VAR	0	0	455	13.45611	.	.	
17	Efficiency		VAR	0	0	100	66.28436	.	.	
18	FG		VAR	0	0.470666	23.4246	1.081501	.	.	
19	IBU		VAR	0	0	3409.3	44.47322	.	.	
20	OG		VAR	0	1	31.7908	1.409707	.	.	
21	Size		VAR	0	1.047	5000	43.65764	.	.	
22	StyleID		VAR	0	1	450	60.05694	.	.	
23	UserId		VAR	68.075	49	134362	43304.31	.	.	

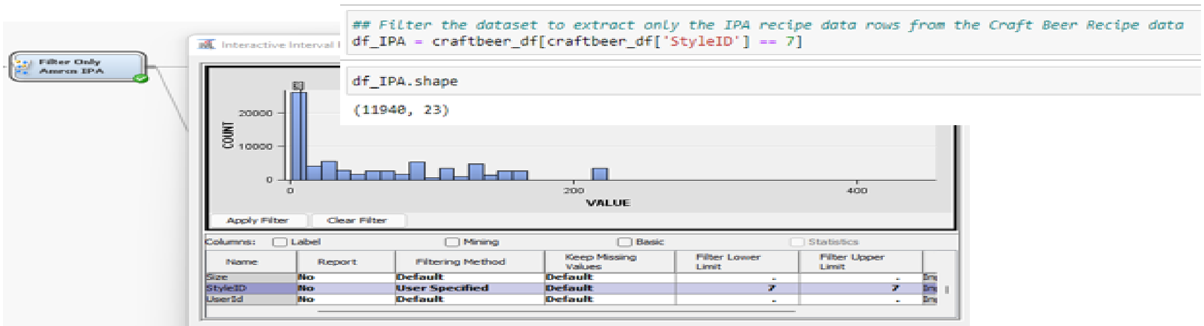
Fig 2. SAS EM – EXPLORE View of Dataset Attributes

Looking at the statistics on the homebrew dataset, it does look like data preparation will be required before we attempt to identify clusters out of the data.

Focus on American IPA First

This task assignment is only interested in American IPA recipes. Although, at 16% of the dataset, American IPA is the single largest style there are **175** other styles included, such as Belgian Blond Ale, Oatmeal Stout and so on.

A quick filter in SAS EM / Python creates an American IPA dataset, which is identified by a ‘*Style_Id*’ = 7.



```
## Filter the dataset to extract only the IPA recipe data rows from the Craft Beer Recipe data
df_IPA = craftbeer_df[craftbeer_df['StyleID'] == 7]
```

```
df_IPA.shape
```

```
(11940, 23)
```

Name	Report	Filtering Method	Keep Missing Values	Filter Lower Limit	Filter Upper Limit
Size	No	Default	Default	.	.
StyleID	No	User Specified	Default	7	7
UserId	No	Default	Default	.	.

Fig 3. SAS EM – Filter Only on America IPA

I chose to do this filtering step before any other data analysis and preparation as I am not interested in cleaning up outliers, missing data, or errors for non-American IPA rows.

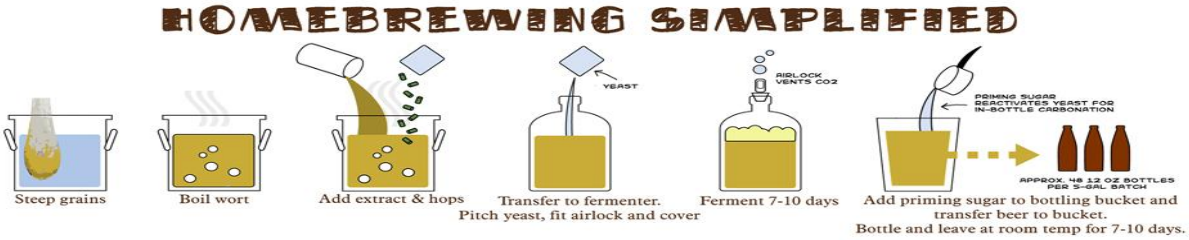
Closer Look at Data Quality

Beer recipe records are user reported in through the Brewer’s Friend website and the quality of numerical data appears to be very good, possibly encouraged by the layout of the data entry webpage. There are some data ranges that look a little suspect, but we will review these specifically in the next section.

The categorical attributes are of a very variable quality. SAS EM reports that there are no missing categorical rows, but it can be seen in Fig. 2 that the most common value for most categorical attributes is ‘N/A’. These attributes largely describe post fermentation activity, and I will return to them in the final stages of this task.

What Attributes are Important for this Clustering Task?

Taking the personal preferences for American IPA into account, as described in Section 1, and looking at this very simple diagram of the homebrew process (below), we can identify the key numerical attributes upon which our cluster analysis should be built.



HOMEBREWING SIMPLIFIED
Steep grains
Boil wort
Add extract & hops
YEAST
Transfer to fermenter. Pitch yeast, fit airlock and cover
AIRLOCK VENTS CO2
Ferment 7-10 days
PRIMING SUGAR REACTIVATES YEAST FOR IN-BOTTLE CARBONATION
APPROX. 48 12 OZ BOTTLES PER 5-GAL BATCH
Add priming sugar to bottling bucket and transfer beer to bucket. Bottle and leave at room temp for 7-10 days.

Fig 4. Homebrew Process[1].

- *OG* - The original gravity (sugar content) of the beer post Wort cooldown.
- *FG* - The final gravity (remaining sugar content) of the beer after fermentation.
- *ABV* - Calculated alcohol by volume, determined by difference between OG and FG.
- *IBU* - International Bittering Units, which is how perceptively bitter the beer is.
- *Size* – Amount (in litres) brewed for specific recipe.
- *Colour* – Light to Dark (zero to 40+ scale).
- *Boil Time* - how long the wort was boiled.
- *Efficiency* - how much possible sugars were extracted from the grains.

These of attributes match the general selection used in other clustering Notebooks on Kaggle**[2]**.

Brief Analysis of Key Clustering Numerical Attributes

The filtering of the homebrew data to only American IPA has removed several of the more obvious outliers and suspicious data elements, such as beers with ABV values between and very unhealthy 40% - 80%, and bitterness levels at an impossible 1000+ score.

However, there are still a range of changes to make to these features to remove certain skews in the data and to fit within the objectives of this task. These changes are elaborated in the next section.

3. Data Preparation

There is no missing data from our required numerical columns in the American IPA sub-dataset, so there is no need to impute or remove rows because of data gaps.

However, using a mixture of domain knowledge and personal preference a certain number of rows will be eliminated based on the following criteria;

- Colour less than 0.5 on the beer colour scale. These rows also correspond to zero/near zero IBU entries. This is practically just water, and presumably an error.

- Efficiency levels above 85%. Values near 100% seem unrealistic for my set-up.

- No recipes aimed at homebrew output greater than 50 litres. This task is not focusing on recipe data for home brew produced at a near industrial levels. There are also some Size values more than 1000 litres that are skewing this data attribute badly.

- IBU values greater than 150. This seems a reasonable threshold in terms of taste but there are also a small range of values stretching from 200 to approximately 1250 that are almost certainly bad data entries and are incorrectly skewing this data element.

Looking at the histograms after the above changes gives us a much more satisfactory set of data elements with which to proceed to the clustering analysis phase of this task.

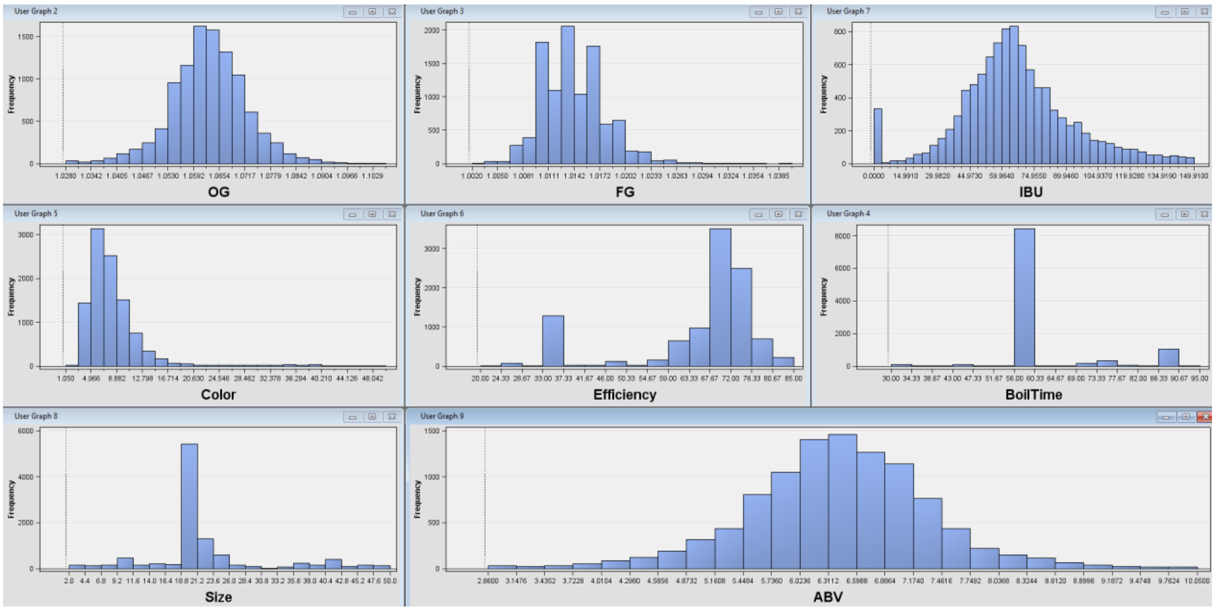


Fig 5. SAS EM Post-Data Preparation View of Key Numerical Attributes

This filtering has also ensured that:

- The OG value is less than the desired upper limit of 1.15
- The FG value is always just above 1.00

The OG and FG data points above, along with many of the other numerical attributes, now conform with data range values that an experienced homebrewer would expect to see in a recipe[2].

4. Details of Algorithms & Configurations

Additional Preparation for Cluster Analysis

Setting up SAS EM for Clustering analysis is relatively straightforward.

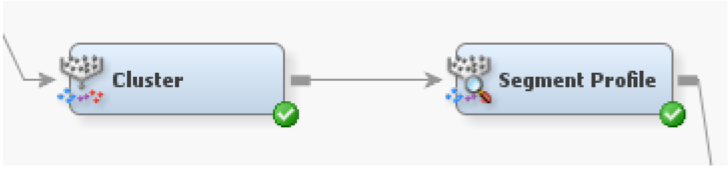
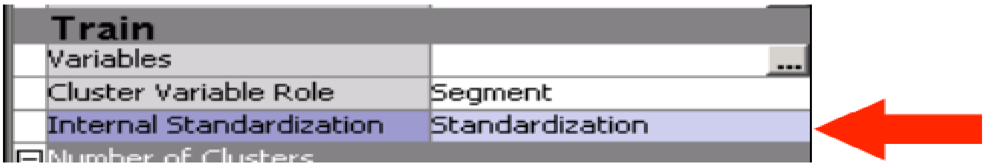


Fig 6. SAS EM Cluster Node Set up

Although it can more correctly be considered part of the Data Preparation process, the numerical data needs to be standardised first before the Cluster Analysis. This is done because for inputs to create clusters they should have similar measurement scales.

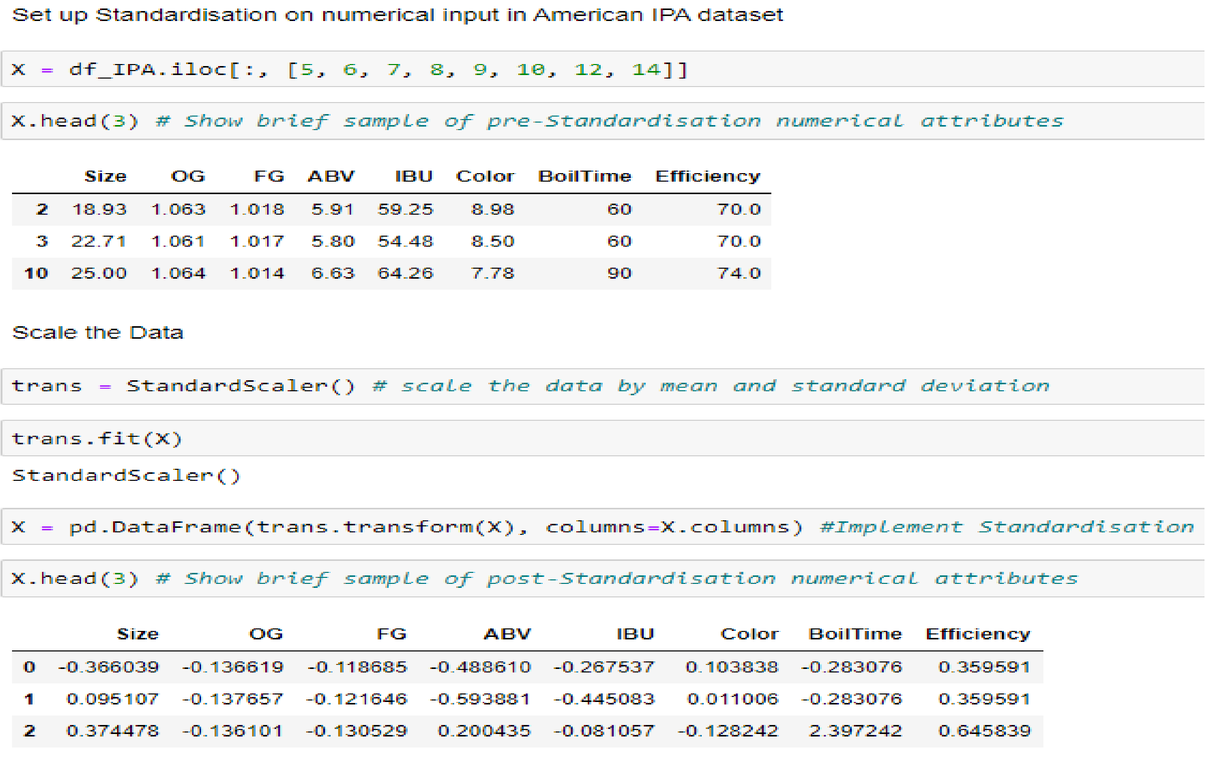
In SAS EM it is a simple setting on the 'Cluster' node:



Train	
Variables	...
Cluster Variable Role	Segment
Internal Standardization	Standardization
Number of Clusters	

Fig 7. SAS EM Cluster Node Setting for Standardisation

In Python it is a few lines of code. A small segment of data is displayed in this diagram to show the effect of the scaling routines.



Set up Standardisation on numerical input in American IPA dataset

```
X = df_IPA.iloc[:, [5, 6, 7, 8, 9, 10, 12, 14]]
```

```
X.head(3) # Show brief sample of pre-Standardisation numerical attributes
```

```
     Size     OG     FG   ABV    IBU  Color  BoilTime  Efficiency
 2  18.93  1.063  1.018  5.91  59.25   8.98        60        70.0
 3  22.71  1.061  1.017  5.80  54.48   8.50        60        70.0
10  25.00  1.064  1.014  6.63  64.26   7.78        90        74.0
```

Scale the Data

```
trans = StandardScaler() # scale the data by mean and standard deviation
```

```
trans.fit(X)
```

```
StandardScaler()
```

```
X = pd.DataFrame(trans.transform(X), columns=X.columns) #Implement Standardisation
```

```
X.head(3) # Show brief sample of post-Standardisation numerical attributes
```

```
       Size        OG        FG       ABV       IBU     Color  BoilTime  Efficiency
0 -0.366039 -0.136619 -0.118685 -0.488610 -0.267537  0.103838 -0.283076    0.359591
1  0.095107 -0.137657 -0.121646 -0.593881 -0.445083  0.011006 -0.283076    0.359591
2  0.374478 -0.136101 -0.130529  0.200435 -0.081057 -0.128242  2.397242    0.645839
```

Fig 8. Python Code for Standardisation

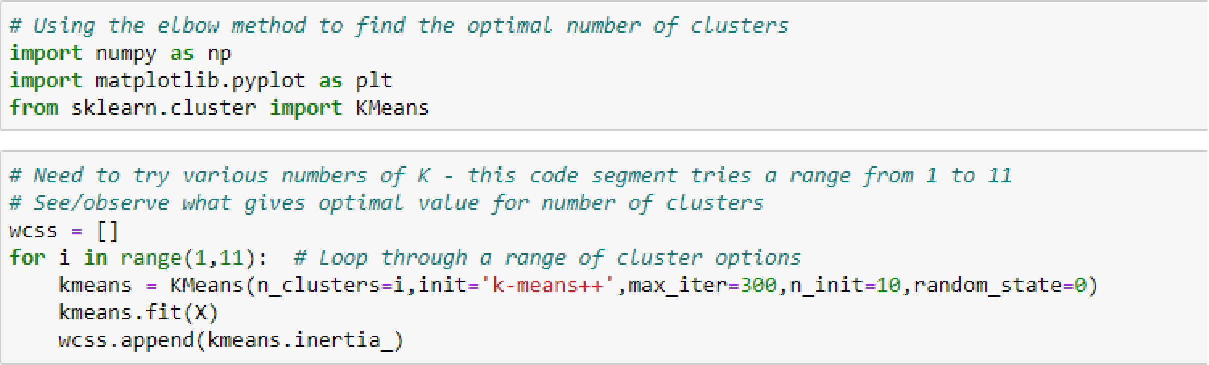
How Many Clusters?

Running the SAS EM *Cluster* node with the default 'Automatic' Specification Method for 'Numbers of Clusters' generates nearly **50** clusters in the node results. In practice. this is an unwieldy number with which to work and process into Segment Profiling of the clusters.

The Python code for cluster analysis is being run in parallel to provide us with additional options to determine a logical number of clusters. A scaled dataset has been created in our Python environment and we can feed this into a KMeans algorithm to determine an optimal number of clusters to use in our homebrew analysis.

The Python code sets up an iteration to generate a graph to which we can apply the 'Elbow Method' to visually assess the appropriate numbers of clusters we should use.

Do the clustering – But how many Clusters – Use the elbow method to determine the optional number of clusters to use

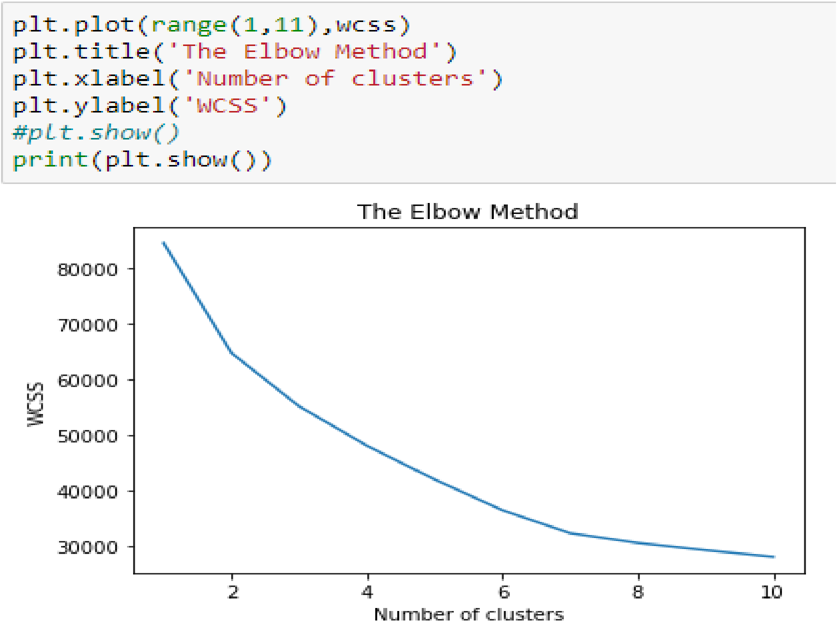


```
# Using the elbow method to find the optimal number of clusters
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
```

```
# Need to try various numbers of K - this code segment tries a range from 1 to 11
# See/observe what gives optimal value for number of clusters
wcss = []
for i in range(1,11):  # Loop through a range of cluster options
    kmeans = KMeans(n_clusters=i,init='k-means++',max_iter=300,n_init=10,random_state=0)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)
```

Fig 9. Python Code for KMeans

The following graph is generated:



```
plt.plot(range(1,11),wcss)
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
#plt.show()
print(plt.show())
```

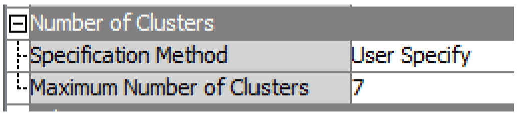
Fig 10. The Elbow Method Graph

In this diagram we can see the elbow bend between 6 and 8. Thus, the optimal cluster number appears to be **7**.

5. Model Performance Metrics & Evaluation of Results

Adjusting the SAS EM Cluster Node Setting and Reviewing Results

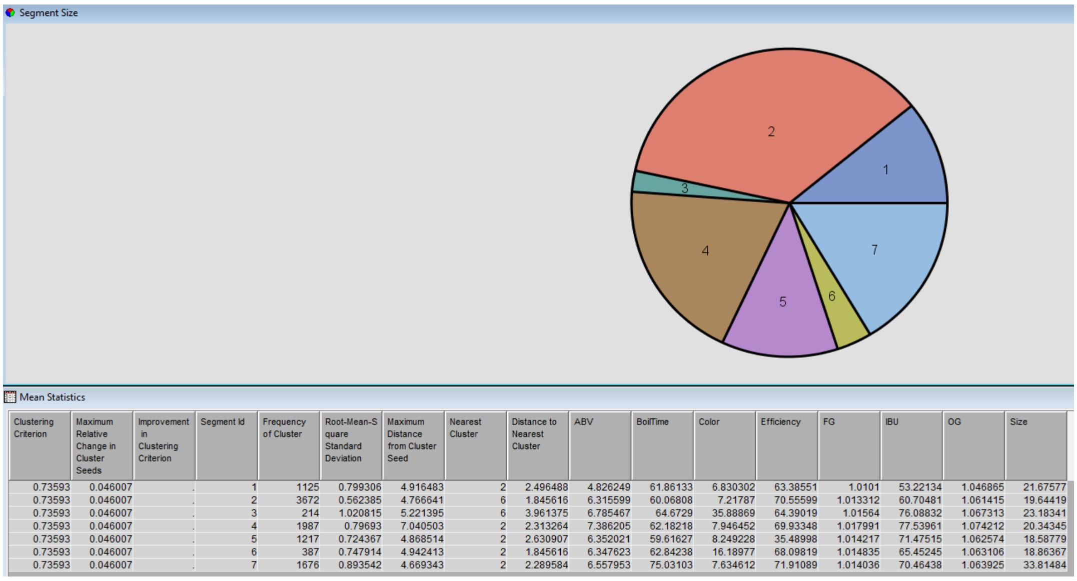
The *Cluster* node in SAS EM allows us to manually set the number of clusters. We will enter '7' based on the Python output from the previous section.



Number of Clusters	
Specification Method	User Specify
Maximum Number of Clusters	7

Fig 11. SAS EM Cluster Node Setting for User Set Cluster Number

The ‘Results’ output from the *Cluster* node represents the 7 clusters statistically and in a pie chart.



Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	ABV	BoilTime	Color	Efficiency	FG	IBU	OG	Size
0.73593	0.046007	.	1	1125	0.799306	4.916483	2	2.496488	4.826249	61.86133	6.830302	63.38551	1.0101	53.22134	1.046865	21.67577
0.73593	0.046007	.	2	3672	0.562385	4.766641	6	1.845616	6.315599	60.06808	7.21787	70.55599	1.013312	60.70481	1.061415	19.64419
0.73593	0.046007	.	3	214	1.020815	5.221395	6	3.961375	6.785467	64.6729	35.88869	64.39019	1.01564	76.08832	1.067313	23.18341
0.73593	0.046007	.	4	1987	0.79693	7.040503	2	2.313264	7.386205	62.18218	7.946452	69.93348	1.017991	77.53961	1.074212	20.34345
0.73593	0.046007	.	5	1217	0.724367	4.868514	2	2.630907	6.352021	59.61627	8.249228	35.48998	1.014217	71.47515	1.062574	18.58779
0.73593	0.046007	.	6	387	0.747914	4.942413	2	1.845616	6.347623	62.84238	16.18977	68.09819	1.014835	65.45245	1.063106	18.86367
0.73593	0.046007	.	7	1676	0.893542	4.669343	2	2.289584	6.557953	75.03103	7.634612	71.91089	1.014036	70.46438	1.063925	33.81484

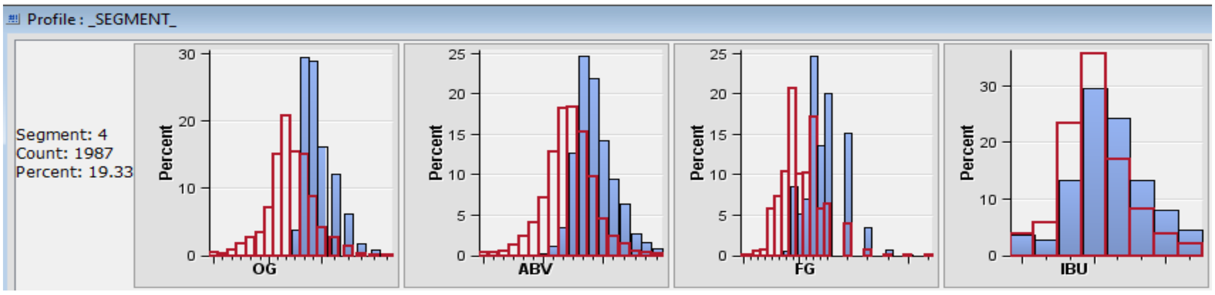
Fig 12. SAS EM Cluster Node Setting for User Set Cluster Number

Although this gives us a high-level view of the Cluster breakdown, it is necessary to proceed to the *Segment Profile* node to gain a better understanding of how the clusters have been determined.

Segment Profiles of Interest

Looking at the results from the *Segment Profile* node we can get a sense of how and why data elements are grouped in each cluster.

For this analysis, the cluster closest to our desired attributes is Cluster **4**. This determination is made based on the graphical results output by the *Segment Profile* node.


Profile : _SEGMENT_

Segment: 4
Count: 1987
Percent: 19.33

Fig 13. SAS EM Cluster 4 Segment Profile

This red overlay shows the average distribution for the shown attributes. This cluster skews slightly higher on ABV and IBU.

Student: Ciaran Finnegan No: D21124026 Prog: TU060 – Part Time (First Year)

In addition, looking at the 'Variable Worth' graph for Cluster 4 it is evident that ABV and IBU are noticeable characteristics in this cluster.

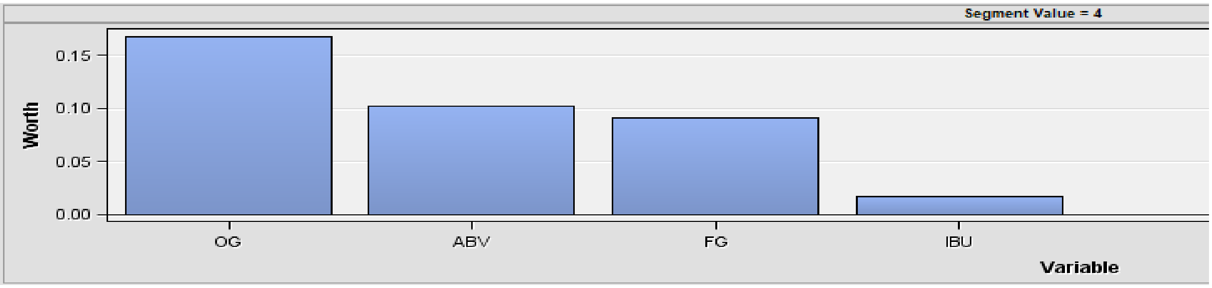


Fig 14. SAS EM Cluster 4 Segment Profile

Final Step: Extracting Usable Cluster 4 Recipes

The *Segment Profile* node labels each dataset row according to its cluster value. The SAS EM diagram has been extended with additional filter nodes to isolate Cluster 4 entries, and then remove any rows that have missing categorical attributes describing the brewing process. Thus, the process leaves us with rows from Cluster 4 that have complete brewing instructions.

This final subset of data is downloaded into an XL spreadsheet to provide the basis for further practical homebrew experimentation.

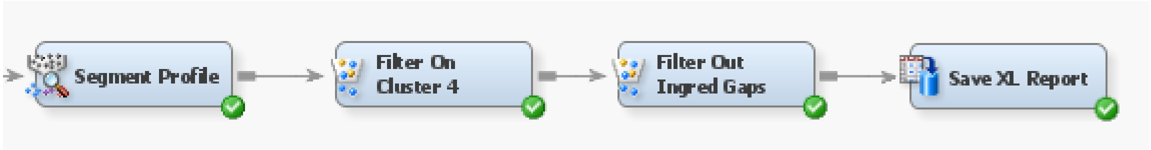


Fig 15. Extracting Clustered Data into EXCEL

(A copy of this EXCEL sheet accompanies this assignment submission).

6. Comparison with other Research & Reflections

1 – Feature Engineering with Domain Knowledge

The Kaggle Notebook by Dereck Bearsong[3], which conducted extensive visualization and analysis of the homebrew recipes was essential to identify erroneous data ranges. This work was informed by a significant amount of domain knowledge and helped pick out values that were skewing inputs to the clustering process. These are discussed in more detail in Section 3 of this task report.

This highlights the advantage of having access to practical experience that can direct feature engineering/data preparation in the machine learning process.

2 – AI and Craft Beer Recipes

It was difficult to find specific research beyond Kaggle on the Brewers Friend dataset, but ResearchGate provided access to an article published this year (2021) on AI-techniques to develop machine learning-built IPA recipe templates[4]. The source data for this work was also a 70K+ dataset of publicly available craft beer homebrew recipes, which then focused on IPA beers. The concepts were in this research were somewhat challenging to understand, involving seven transformer networks being trained on an end-to-end brew process.

Although I did not take any specific learning from this article into my direct analysis, it gave me confidence that my dataset had significant future AI-driven homebrew potential.

3 – Malt and Hops: Too many variations.

A recent study in 2020 by the UCD Geary Institute (Ireland) drew from the Brewers Friend dataset and conducted an analysis of the impact of regional ingredients on beers[5]. However, a key difference in the UCD analysis as compared to the Kaggle dataset, is that their scraping of the website also included data on malt and hops content.

An interesting point that this study highlighted with this malt/hops data is that the 70K+ recipes listed in the dataset described names for 4,882 different malts, and 5023 separate Hops. Much of this variation was down to regional naming conventions. Upon further analysis, the dataset has only 170 and 229 genuinely unique malts and hops respectively. This had a material impact on the UCD conclusions around the impact of regional ingredients.

If I were to repeat my American IPA analysis, I would investigate the possibility of scrapping the additional malt and hops values from the Brewers Friend website, and 'normalise' the data into a smaller subset of unique numerical values (as in the UCD study). It would be interesting to see if this impacted on the outcome from the IPA Clustering analysis.

7. References

[1] Lachance, S., 2017. *Beer Recipe Exploratory Analysis*. [online] Kaggle.com. Available at: <https://www.kaggle.com/samlac79/beer-recipe-exploratory-analysis> [Accessed 14 December 2021].

[2] Allen, A., 2020. *Beer Embedding Visualization*. [online] Kaggle.com. Available at: <https://www.kaggle.com/volperosso/beer-embedding-visualization> [Accessed 13 December 2021].

[3] Bearsong, D., 2018. *Quick visualization & analysis of Homebrew Recipes*. [online] Kaggle.com. Available at: <https://www.kaggle.com/blasterbrewmaster/quick-visualization-analysis-of-homebrew-recipes> [Accessed 13 December 2021].

[4] Bravin, M., Pfaffli, D., Kuhn, K. and Pouly, M., 2021. Towards Crafting Beer with Artificial Intelligence. *2021 8th Swiss Conference on Data Science (SDS)*, [online] pp.54-55. Available at: <https://ieeexplore.ieee.org/document/9474588> [Accessed 13 December 2021].

[5] Buarque, B., Davies, R., Hynes, R. and Kogler, D., 2020. *Hops, skip & a Jump: The regional uniqueness of beer styles*. [online] Hdl.handle.net. Available at: <http://hdl.handle.net/10419/237578> [Accessed 13 December 2021].

TASK 2 – *Text Mining – Comparison of Rotten Tomatoes Movie Reviews in Word Clouds*

1. Definition of Problem

Rotten Tomatoes is a very well know movie database website that allows critics and the public (defined as 'Users' in this assignment) to upload reviews and scores on new movie releases.

It has often been claimed that film critics are out of touch with the taste of regular movie goers and that this can be seen by the dichotomy between critic and user scores on Rotten Tomatoes, particularly for the larger commercial ('blockbuster') movie titles.

This assignment attempts to provide some data analytical rigour to this assertion of 'aloofness' by movie critics. Although one can just compare the scores given by critics and users to movies, this assignment attempts to identify noticeable differences in the patterns of the language used in the reviews themselves.

Review data from four of the most recent movies in the Marvel Cinematic Universe is pulled from the Rotten Tomatoes website, and Text Mining techniques are used to present opposing WordClouds built from the critic and user reviews.

Although it is only a sample of four movies, this Word Cloud analyse will provide a visual indication of the disparity, if any, between the reviews from critics and users.

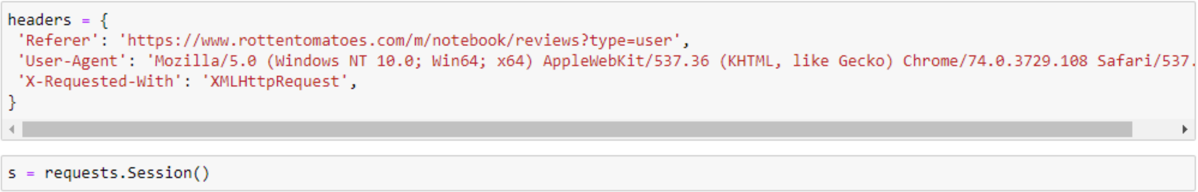
2. Data Exploration & Descriptive Analytics

For simplicity, the assignment focuses on the Word Cloud analysis of the reviews for the Marvel movie 'Eternals' in particular. However, the Python code is built in a generic way to extract and analyse the review data from any given move on the Rotten Tomatoes website.

Calling APIs

For reasons explained in Section 6 of this section of the report, the data extraction from Rotten Tomatoes was carried out via the publicly available API.

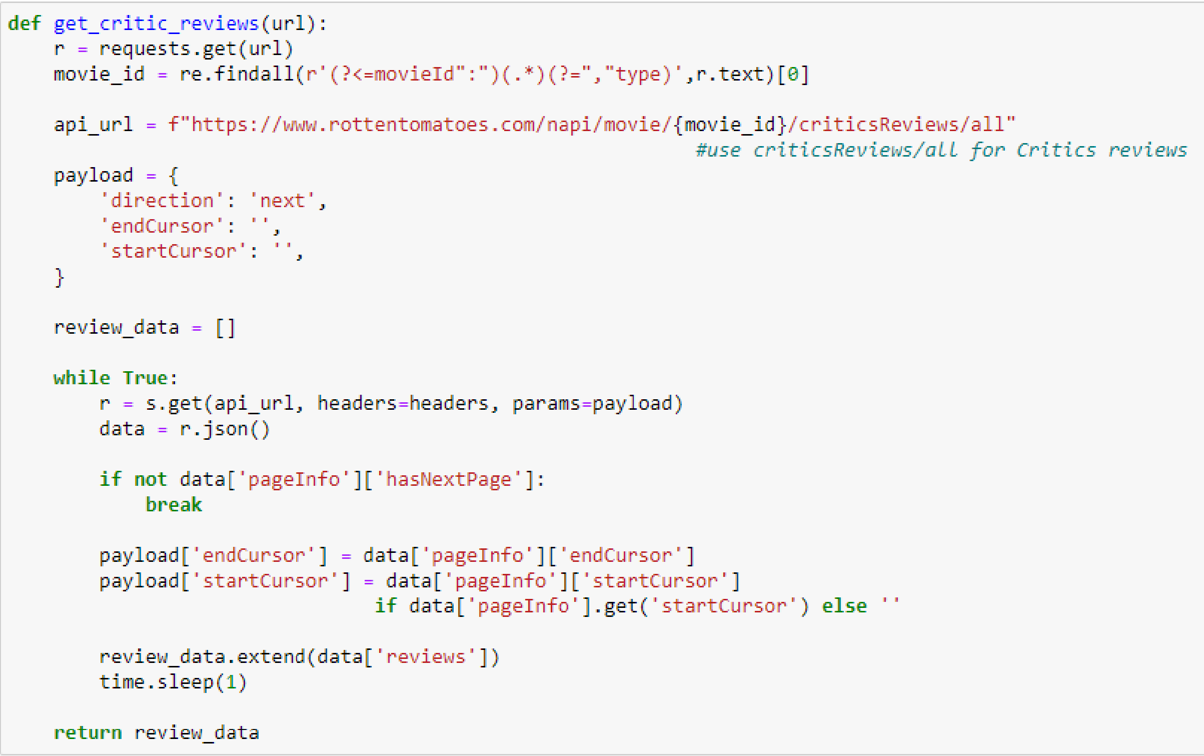
Based on technical specifications and code snippets from a StackOverflow article[1], two functions were written to separately extract the critics and the users movie reviews.



```
headers = {
 'Referer': 'https://www.rottentomatoes.com/m/notebook/reviews?type=user',
 'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/74.0.3729.108 Safari/537.
 'X-Requested-With': 'XMLHttpRequest',
}
```

```
s = requests.Session()
```

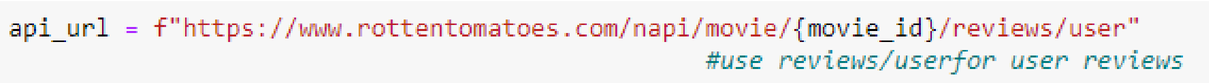
Fig 1 – Setting up 'header' variable for API Call



```
def get_critic_reviews(url):
    r = requests.get(url)
    movie_id = re.findall(r'(?<=movieId":")(.*)(?=","type)',r.text)[0]

    api_url = f"https://www.rottentomatoes.com/napi/movie/{movie_id}/criticsReviews/all"
                                                        #use criticsReviews/all for Critics reviews

    payload = {
        'direction': 'next',
        'endCursor': '',
        'startCursor': '',
    }

    review_data = []

    while True:
        r = s.get(api_url, headers=headers, params=payload)
        data = r.json()

        if not data['pageInfo']['hasNextPage']:
            break

        payload['endCursor'] = data['pageInfo']['endCursor']
        payload['startCursor'] = data['pageInfo']['startCursor']
                                if data['pageInfo'].get('startCursor') else ''

        review_data.extend(data['reviews'])
        time.sleep(1)

    return review_data
```

Fig 2 – Python Function to Extract Movie Reviews from Critics for ‘Eternals’

The function to extract the User reviews differs only in the API line.



```
api_url = f"https://www.rottentomatoes.com/napi/movie/{movie_id}/reviews/user"
                                                        #use reviews/userfor user reviews
```

Fig 3 – Code Snippet to Invoke APIS for User Reviews

I borrowed heavily from publicly available code to call the API call because of the challenges of web scrapping that were impeding my progress on the assignment.

Obviously, it would be better programming practice to have a single function to extract reviews and parse the API call using a flag in the function parameter. For several reasons, I kept the functions separate, despite the duplication in code:

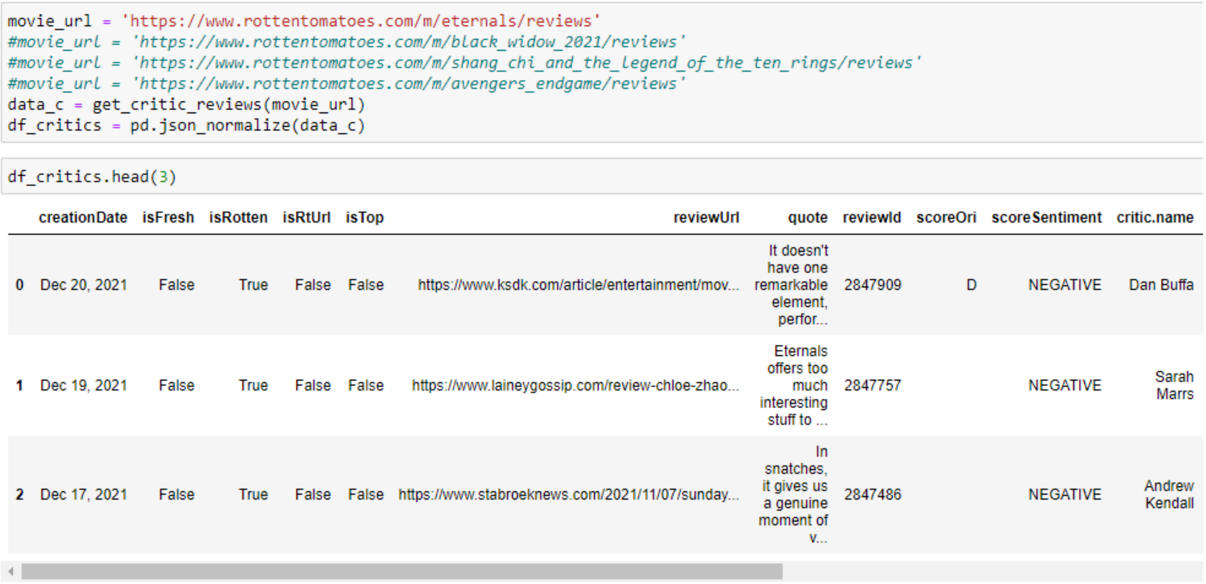
- Simplicity. The primary focus of this assignment is intended to be on setting optimal parameters for the Word Cloud display and I wanted to move quickly to that stage of development.

- The User reviews are considerably larger in volume than the review data from the Critics and take a significant amount of time to extract, usually in the order of several minutes. I experimented with different ‘Sleep()’ values for both sets of review for ‘Eternals’.

Working with Dataframes

The output from the APIs populates dataframes with a collection of data attributes for a given movie.

The dataframes returned by the APIs differ in structure between Critics and Users.



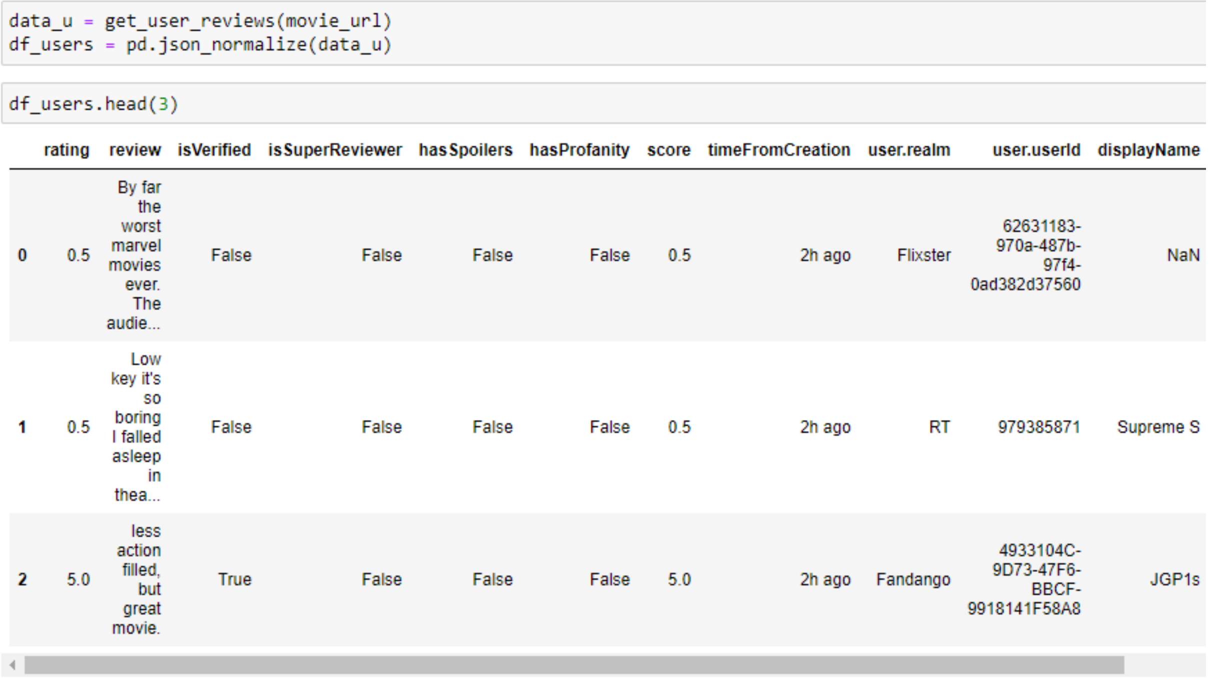
```
movie_url = 'https://www.rottentomatoes.com/m/eternals/reviews'
#movie_url = 'https://www.rottentomatoes.com/m/black_widow_2021/reviews'
#movie_url = 'https://www.rottentomatoes.com/m/shang_chi_and_the_legend_of_the_ten_rings/reviews'
#movie_url = 'https://www.rottentomatoes.com/m/avengers_endgame/reviews'
data_c = get_critic_reviews(movie_url)
df_critics = pd.json_normalize(data_c)
```

```
df_critics.head(3)
```

	creationDate	isFresh	isRotten	isRtUrl	isTop	reviewUrl	quote	reviewId	scoreOri	scoreSentiment	critic.name
0	Dec 20, 2021	False	True	False	False	https://www.ksdk.com/article/entertainment/mov...	It doesn't have one remarkable element, perfor...	2847909	D	NEGATIVE	Dan Buffa
1	Dec 19, 2021	False	True	False	False	https://www.laineygossip.com/review-chloe-zhao...	Eternals offers too much interesting stuff to ...	2847757		NEGATIVE	Sarah Marrs
2	Dec 17, 2021	False	True	False	False	https://www.stabroeknews.com/2021/11/07/sunday...	In snatches, it gives us a genuine moment of v...	2847486		NEGATIVE	Andrew Kendall

Fig 4 – Dataframe for Critics Reviews

Better Python coding practice would have been to set up a dictionary of the movie URLs and run the process through a loop to generate WordClouds. For simplicity in this assignment, I just created one Jupyter Notebook per movie.



```
data_u = get_user_reviews(movie_url)
df_users = pd.json_normalize(data_u)
```

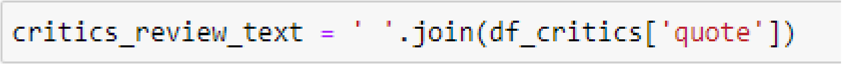
```
df_users.head(3)
```

	rating	review	isVerified	isSuperReviewer	hasSpoilers	hasProfanity	score	timeFromCreation	user.realm	user.userId	displayName
0	0.5	By far the worst marvel movies ever. The audie...	False	False	False	False	0.5	2h ago	Flixster	62631183-970a-487b-97f4-0ad382d37560	NaN
1	0.5	Low key it's so boring I falled asleep in thea...	False	False	False	False	0.5	2h ago	RT	979385871	Supreme S
2	5.0	less action filled, but great movie.	True	False	False	False	5.0	2h ago	Fandango	4933104C-9D73-47F6-BBCF-9918141F58A8	JGP1s

Fig 5 – Dataframe for User Reviews

Running a simple line of code pulls out and concatenates all the review text from each dataframe.

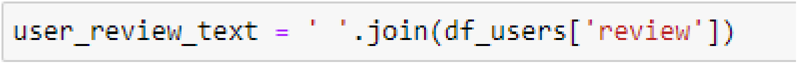
The ‘quote’ column contains the text from Critic reviews for ‘Eternals’.



```
critics_review_text = ' '.join(df_critics['quote'])
```

Fig 6 – Extract and Concatenate Critic Reviews into Text Variable

The ‘review’ column contains the text from User reviews for ‘Eternals’.



```
user_review_text = ' '.join(df_users['review'])
```

Fig 7 – Extract and Concatenate User Reviews into Text Variable

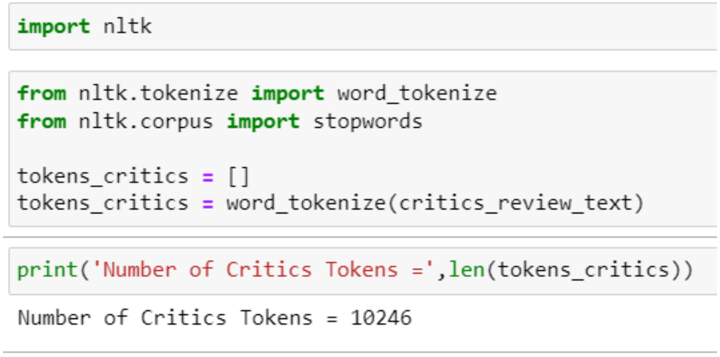
These text variables then form the starting point of the data preparation phase in the following section.

3. Data Preparation

NLTK, or Natural Language Toolkit, is a Python package that we will use to carry out **text pre-processing** tasks before creating the WordClouds.

Tokenize the Words

The first step in the data preparation process is to Tokenize the text. This begins the process of turning unstructured data into structured data, which will be easier to analyse and present in WordClouds. In our Python code we begin by breaking the critic and user reviews into individual words.

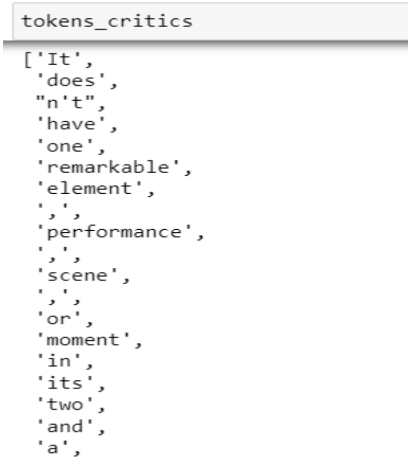


```
import nltk
```

```
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords

tokens_critics = []
tokens_critics = word_tokenize(critics_review_text)
```

```
print('Number of Critics Tokens =',len(tokens_critics))

Number of Critics Tokens = 10246
```

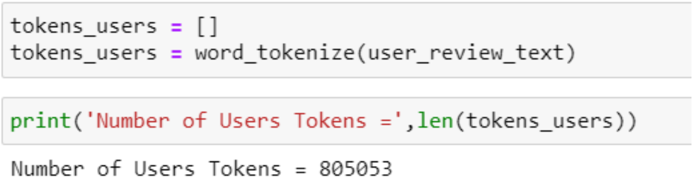
Fig 8 – Critic Reviews Tokenized



```
tokens_critics
```

```
['It',
 'does',
 "n't",
 'have',
 'one',
 'remarkable',
 'element',
 ',',
 'performance',
 ',',
 'scene',
 ',',
 'or',
 'moment',
 'in',
 'its',
 'two',
 'and',
 'a',
```

Fig 9 – Partial View of Tokens for Critic Reviews



```
tokens_users = []
tokens_users = word_tokenize(user_review_text)
```

```
print('Number of Users Tokens =',len(tokens_users))
```

```
Number of Users Tokens = 805053
```

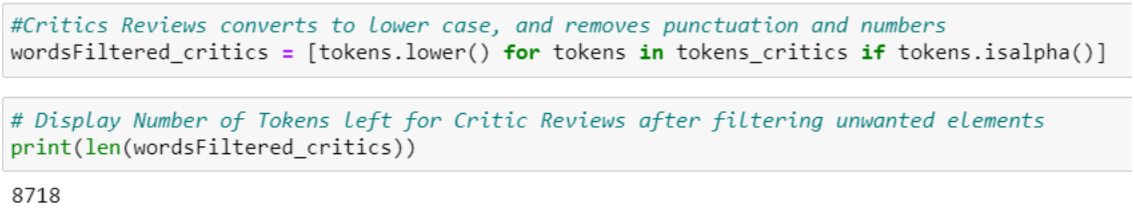
Fig 10 – User Reviews Tokenized

Note, the significantly greater volume of User reviews, which took many minutes to extract.

Remove Numbers and Punctuation: Filter Tokens

The text of the reviews only needs to contain words. Numbers and punction will be of no value in the analysis we present in the WordClouds.

The following code will filter the tokens and eliminate the text that is not required. It is also good practice to reduce all characters to lower case.

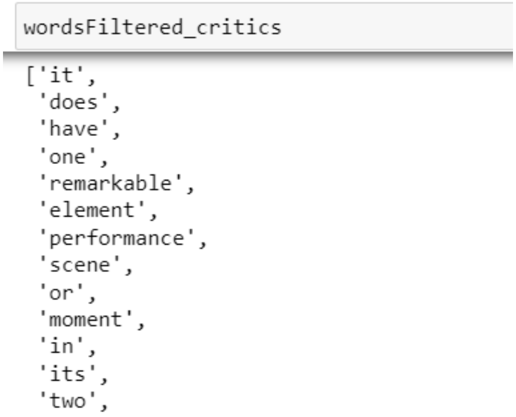


```
#Critics Reviews converts to lower case, and removes punctuation and numbers
wordsFiltered_critics = [tokens.lower() for tokens in tokens_critics if tokens.isalpha()]
```

```
# Display Number of Tokens left for Critic Reviews after filtering unwanted elements
print(len(wordsFiltered_critics))
```

```
8718
```

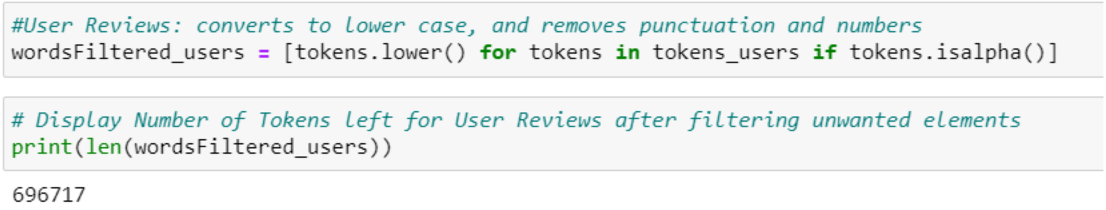
Fig 11 – Critic Reviews Filtered for Punctuation and Numbers



```
wordsFiltered_critics
```

```
['it',
 'does',
 'have',
 'one',
 'remarkable',
 'element',
 'performance',
 'scene',
 'or',
 'moment',
 'in',
 'its',
 'two',
```

Fig 12 – Critic Reviews – Filtered (sample)



```
#User Reviews: converts to lower case, and removes punctuation and numbers
wordsFiltered_users = [tokens.lower() for tokens in tokens_users if tokens.isalpha()]
```

```
# Display Number of Tokens left for User Reviews after filtering unwanted elements
print(len(wordsFiltered_users))
```

```
696717
```

Fig 13 – User Reviews Filtered for Punctuation and Numbers

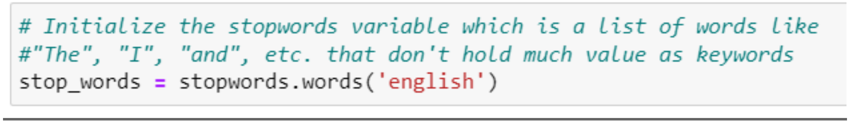
Again, we significant disparity in word volume.

Setup Stop Words

Stop words are words that should be ignored and will be filtered out in this stage of text pre-processing.

Very common words like `'in'`, `'is'`, and `'an'` are often used as stop words as they offer little meaning to a text in and of themselves.

The Python NLTK library contains several pre-defined English stop words, which shall be removed first.

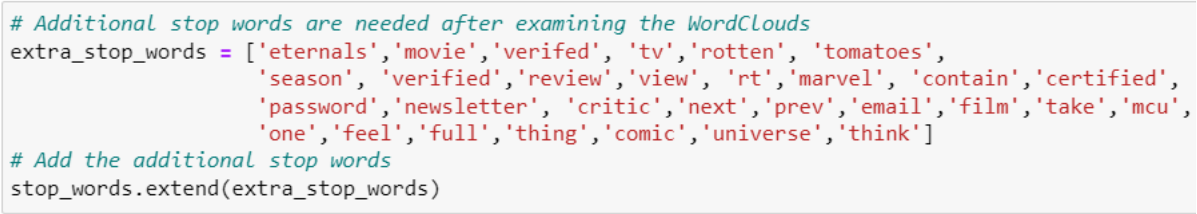


```
# Initialize the stopwords variable which is a list of words like
#"The", "I", "and", etc. that don't hold much value as keywords
stop_words = stopwords.words('english')
```

Fig 14 – Set Up Standard English ‘Stop Words’

The quality of the analysis presented by the WordClouds will be improved by adding additional stop words to the list. In this assignment, as the analysis focused on Marvel movies, both sets of reviewers would use words like *‘mcu’*, *’film’*, *’universe’*, *’comic’* and so on that did not provide useful context or differentiation. The title of the film was also the most common phrase in both sets of reviews so that was included in the stop word list.

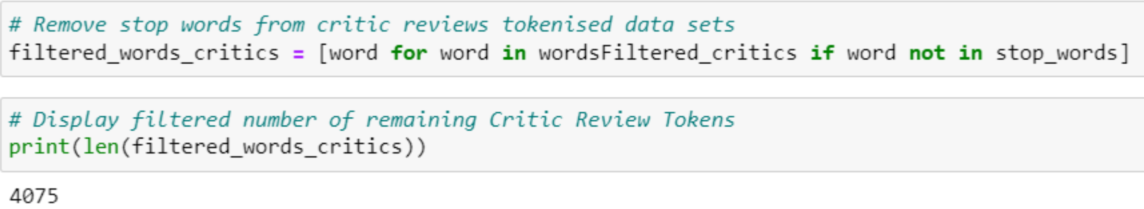
The final list of custom stop words was arrived after an iteration process that generated WordClouds and identified words and phrase that added little value to the data visualisation.



```
# Additional stop words are needed after examining the WordClouds
extra_stop_words = ['eternals','movie','verifed', 'tv','rotten', 'tomatoes',
                    'season', 'verified','review','view', 'rt','marvel', 'contain','certified',
                    'password','newsletter', 'critic','next','prev','email','film','take','mcu',
                    'one','feel','full','thing','comic','universe','think']
# Add the additional stop words
stop_words.extend(extra_stop_words)
```

Fig 15 – Add Additional Assignment Specific 'Stop Words'

These steps reduce the number of tokens in each set of reviews.

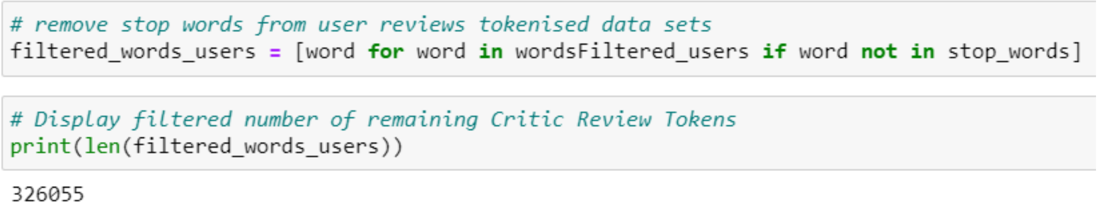


```
# Remove stop words from critic reviews tokenised data sets
filtered_words_critics = [word for word in wordsFiltered_critics if word not in stop_words]
```

```
# Display filtered number of remaining Critic Review Tokens
print(len(filtered_words_critics))
```

```
4075
```

Fig 16 – Filter Stop Words from Critic Review Text: 4075 tokens remain



```
# remove stop words from user reviews tokenised data sets
filtered_words_users = [word for word in wordsFiltered_users if word not in stop_words]
```

```
# Display filtered number of remaining Critic Review Tokens
print(len(filtered_words_users))
```

```
326055
```

Fig 17 – Filter Stop Words from User Review Text: 326055 tokens remain

4. Configuration of WordCloud Parameters

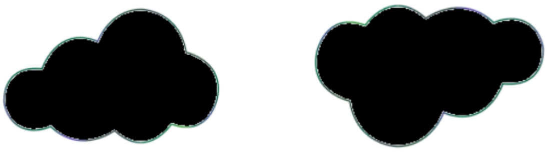
At this point in the assignment the next step is to create a WordCloud for each finalised data set of tokenised words for reviews. The WordCloud Python library will create the WordClouds, which will then be displayed as side-by-side subplots by the matplotlib library.

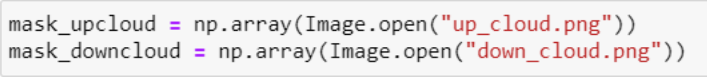
Shaping the WordClouds

To make the WordClouds more visually appealing, the shape of the display is being converted from a square into a cloud outline (the image could also be interpreted as a 'think bubble').

The Critics reviews 'cloud' will be presented in the upper part of the subplot display in black. The User Reviews will render in red using an inverted cloud image as the outline.

To create the 'mask' for the WordCloud function parameter, two *png* images were created and added to the assignment file directory.





```
mask_upcloud = np.array(Image.open("up_cloud.png"))
mask_downcloud = np.array(Image.open("down_cloud.png"))
```

Fig 18 – Outline Images to Shape WordClouds

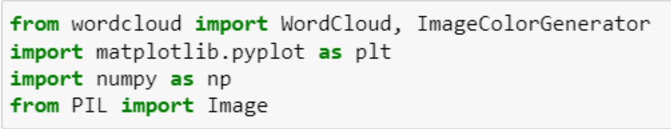
The application of these masks is explained in the following section on WordCloud function parameters.

Configuring the WordClouds

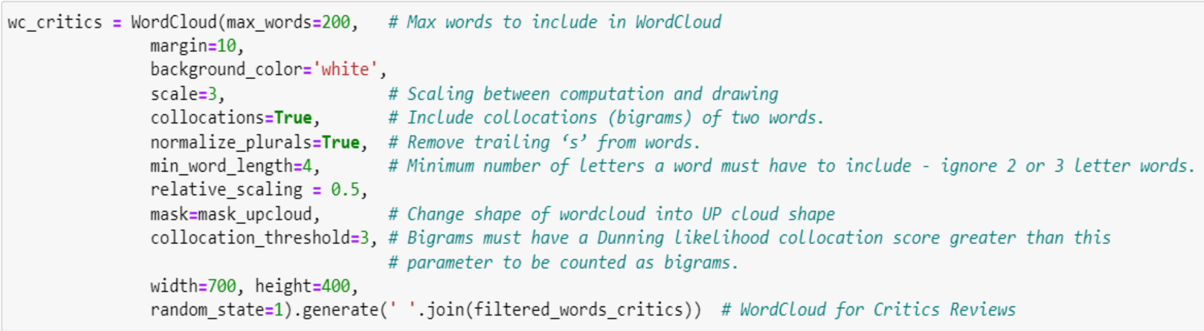
The initial objective of the WordClouds in this assignment was to give visual prominence to words most frequently used by Critics and Users and compare the displays.

Single words did provide some interesting comparisons but additional research on WordClouds generated from the text of speeches in recent American political campaigns[2] indicated that more value might be extracted if short phrases were included. Therefore, the parameters were set on the WordCloud to look at bigrams (tokens more frequently seen in sequence).

The configuration for the Critic Review WordCloud function is given in the following diagram:



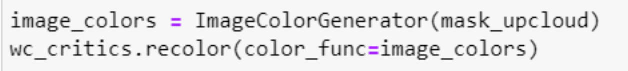
```
from wordcloud import WordCloud, ImageColorGenerator
import matplotlib.pyplot as plt
import numpy as np
from PIL import Image
```



```
wc_critics = WordCloud(max_words=200,   # Max words to include in WordCloud
                       margin=10,
                       background_color='white',
                       scale=3,               # Scaling between computation and drawing
                       collocations=True,     # Include collocations (bigrams) of two words.
                       normalize_plurals=True, # Remove trailing ‘s’ from words.
                       min_word_length=4,     # Minimum number of letters a word must have to include - ignore 2 or 3 letter words.
                       relative_scaling = 0.5,
                       mask=mask_upcloud,     # Change shape of wordcloud into UP cloud shape
                       collocation_threshold=3, # Bigrams must have a Dunning likelihood collocation score greater than this
                                              # parameter to be counted as bigrams.
                       width=700, height=400,
                       random_state=1).generate(' '.join(filtered_words_critics))  # WordCloud for Critics Reviews
```

Fig 19 – Critic Reviews: WordCloud Function Parameters

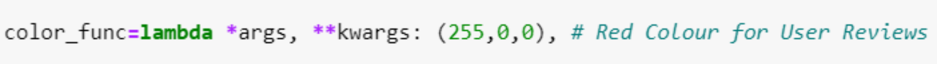
The colour of the Critics review WordCloud is determined by the black colour of the mask used.



```
image_colors = ImageColorGenerator(mask_upcloud)
wc_critics.recolor(color_func=image_colors)
```

Fig 20 – Critic Reviews: Setting the Colour

The parameters for the User reviews WordCloud function are identical, with the exception of another line to alter the colour to ***red***.



```
color_func=lambda *args, **kwargs: (255,0,0), # Red Colour for User Reviews
```

Fig 21 – User Reviews: Setting the Colour

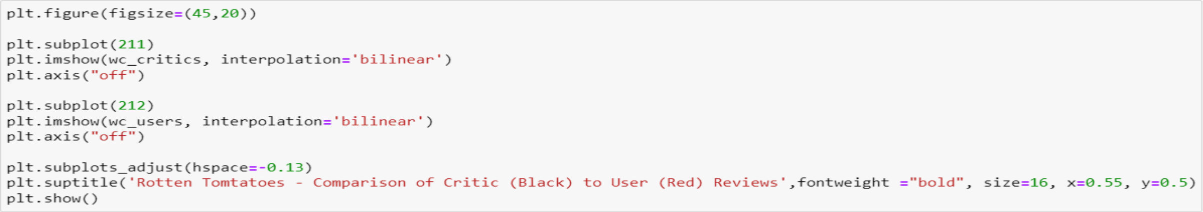
The key WordCloud function parameters, and their purpose in this assignment, can be briefly described as;

- ***collocations***=True : Allows two-word phrases to be displayed. The result of this is evident in the final WordCloud display.
- **min_word_length**=4 : Avoids smaller less meaningful words/phrases
- **collocation_threshold**=3 : This sets a threshold above which a bigram must score to be considered for the WordCloud

The '*collocation_threshold'* was a value and concept that was somewhat challenging to comprehend but, through trial and error, it did help generate interesting visual comparisons in the review WordClouds.

5. WordCloud Displays

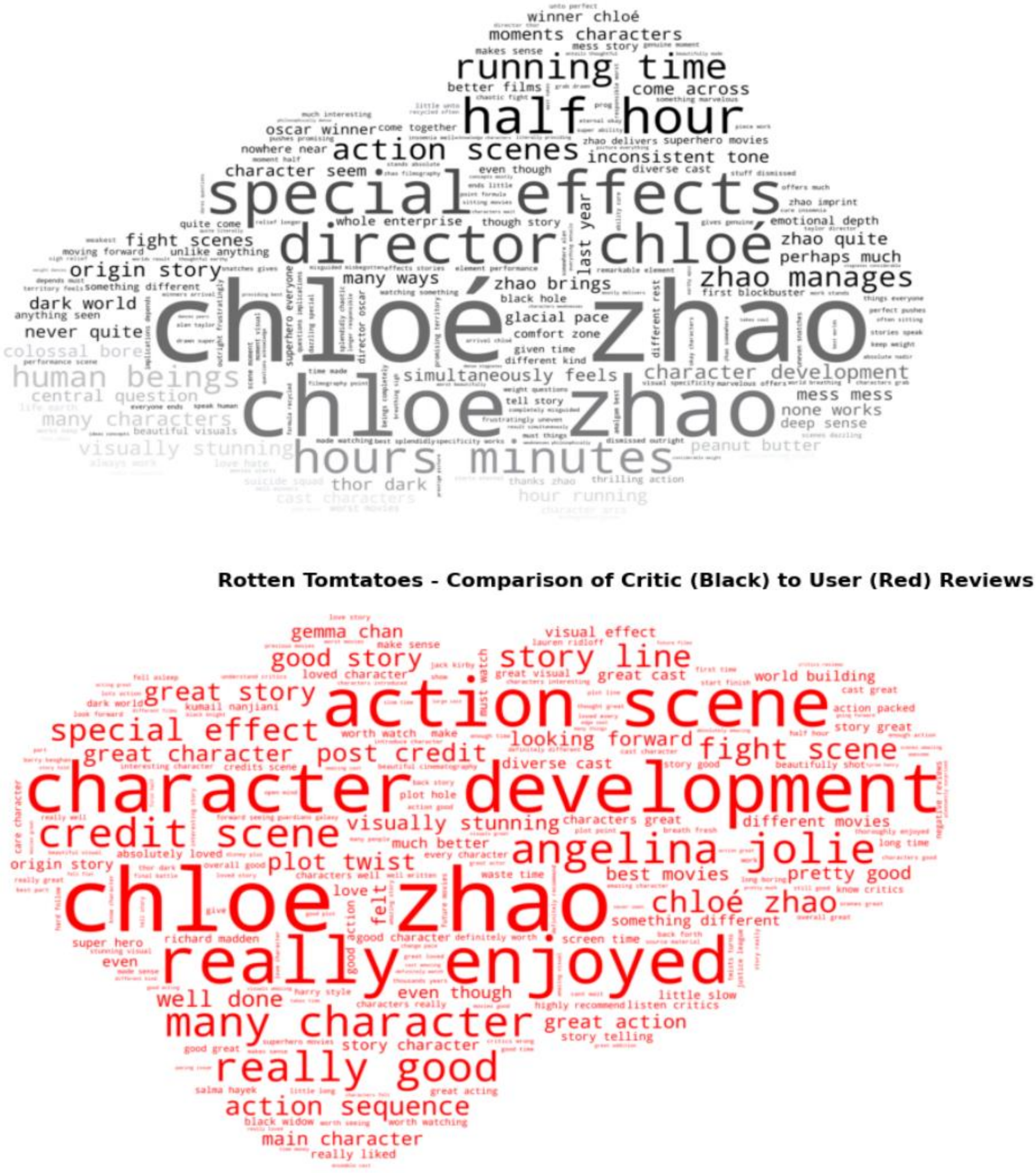
Some simple Python code rendered the WordClouds for Critic and User reviews of the Marvel movie ‘Eternals’.



```
plt.figure(figsize=(45,20))

plt.subplot(211)
plt.imshow(wc_critics, interpolation='bilinear')
plt.axis("off")

plt.subplot(212)
plt.imshow(wc_users, interpolation='bilinear')
plt.axis("off")

plt.subplots_adjust(hspace=-0.13)
plt.suptitle('Rotten Tomtatoes - Comparison of Critic (Black) to User (Red) Reviews',fontweight ="bold", size=16, x=0.55, y=0.5)
plt.show()
```

Fig 22 – Render and Format Position of Review WordClouds

The resulting sub-plots provide interesting insight into the differences between the reviews.



6. Comparison with other Research & Reflections

1 – Presenting Opposing WordClouds

This Text Mining assignment had originally intended to look at the difference in language tone across news websites covering the same, or similar, political events. Although the approach and subject matter of the assignment changed significantly, there were some research papers relating to Text Mining of news outlets that continued to provide guidance on the use and presentation of WordClouds.

Of note was an article written by in 2012 Hensinger, Flaounas and Cristianini at Bristol University entitled *The Appeal of Politics on Online Readers*[3]. The authors employed an opposing set of WordClouds to show what words most and least appealed to readers in terms of their likelihood to read an article in Forbes magazine.

This technique seemed to fit well with the objective of this assignment, and I followed their approach to generate a single contrasting 'at-a-glance' view, from a Rotten Tomatoes movie page, of both the critic and user review WordClouds.

2 – Bigrams and Working with Python WordCloud Parameters

Although the initial WordClouds generated in this assignment were providing a reasonable sense of review content, the focus on displaying single words was removing some valuable context. Looking at related research on sentiment analysis from Rotten Tomatoes there was a reference in the paper from Sorostinean, M., Sana, K., Mohamed, M. and Targhi, A., 2017. *Sentiment Analysis on Movie Reviews*[4] that recommended the use of Bigrams. (A bigram is a sequence of two adjacent elements from a string of tokens.)

Researching the use of Python WordClouds, I found a further article on the *towardsdatascience.com* website entitled *Generate Meaningful Words in Python*[2]. This provided some practical examples on settings for the parameters in a Python WordCloud so that I could extract meaningful short phrases from the reviews, along with single words.

The recommended setting of eliminating words with fewer than four characters and then setting the collocation threshold to '3' produced more meaningful WordClouds. This is described in more detail in Section 4 of this section of the assignment.

I am conscious that I am greatly simplifying the description of the research output to which I referred, and that there was also an element of trial and error in the settings, but I felt the Rotten Tomatoes WordCloud output was very satisfactory.

3 – Calling APIs vs. Web Scrapping

My initial Python code used the *BeautifulSoup* library to scrape the data directly from the Rotten Tomatoes web page for the given Marvel movie.

This has an immediate limitation in that it only returned the text for the reviews visible on the first web page. There were several online resources that provided guidance on URL manipulation as a possible solution. However, the Rotten Tomatoes website is constantly evolving in terms of its structure and the *BeautifulSoup* approach was not proving very robust in the face of these changes.

Looking over online research papers in movie reviews and text mining, I found that others had begun their projects[5] by using the publicly available APIs that websites such as Rotten Tomatoes provide for data retrieval. Following this trail of breadcrumbs led to online technical specifications and sample code that greatly simplified the extraction of Rotten Tomatoes review data.

Reflections

The inclusion of short two-word phrases helps show, through our WordClouds, that critics tend to focus on directors and more technical aspects of the movie, while users are more direct with their emotional enthusiasm.

Our 'Eternals' WordCloud for Users was able to capture phrases like '*really good'* and '*really enjoyed'*. This type of phrase is conspicuous by its absence in the critic reviews, although I did enjoy reading the phrases '*colossal bore'* and '*glacial pace'* buried in the Critics WordCloud.

For the film 'Eternals' the Users were clearly more positive than the Critics, although this pattern is less obvious for the other movies, which can be seen in the attached WordCloud movie images below.

Black Window
Crtics v Users Word(

Shang Chi Ten
Rings Crtics v Users

Endgame Crtics v
Users WordClouds [

For example, Users seem to make more of an issue around the plot of 'Endgame' than was the case with the Critics. (These image files were embedded in the original WORD document. They are included with the PDF for the final assignment submission).

7. References

[1] Adriaansen, R., 2021. *Scraping all reviews of a movie from Rotten Tomato using soup*. [online] Stack Overflow. Available at: <https://stackoverflow.com/questions/69963743/scraping-all-reviews-of-a-movie-from-rotten-tomato-using-soup> [Accessed 20 December 2021].

[2] Dickenson, B., 2020. *Generate Meaningful Word Clouds in Python*. [online] towardsdatascience.com. Available at: <https://towardsdatascience.com/generate-meaningful-word-clouds-in-python-5b85f5668eeb> [Accessed 20 December 2021].

[3] Hensinger, E., Flaounas, I. and Cristianini, N., 2012. *The Appeal of Politics on Online Readers*. [online] Blogs.oii.ox.ac.uk. Available at: <http://blogs.oii.ox.ac.uk/ipp-conference/sites/ipp/files/documents/HensingerFlaounasCristianini_Oxford2012.pdf> [Accessed 20 December 2021].

[4] Sorostinean, M., Sana, K., Mohamed, M. and Targhi, A., 2017. Sentiment Analysis on Movie Reviews. *Journal Agroparistech.*, [online] Available at: <http://www.agroparistech.fr/ufr-info/membres/cornuejols/Teaching/Master-AIC/PROJETS-M2-AIC/PROJETS-2016-2017/main(Amal%20Targhi-%20Mihaela%20SOROSTINEAN-%20Katia%20Sana-Mohamed%20Mohamed).pdf> [Accessed 20 December 2021].

[5] Schaible, J., Carevic, Z., Hopt, O. and Zapilko, B., 2015. *Utilizing the Open Movie Database API for Predicting the Review Class of Movies*. [online] Citeseerx.ist.psu.edu. Available at: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1071.6147&rep=rep1&type=pdf> [Accessed 20 December 2021].

TASK 3 - *Association Rules – Grocery Market Basket Analysis*

1. Definition of Problem

This part of the assignment looks at the concept of association rules being used in market basket analysis.

It begins with a Kaggle sourced dataset (located here) of purchase orders by customers in a grocery store. The objective of this exercise is to use SAS Enterprise Miner (EM) to analyse the grocery order dataset and determine what types of associations and correlations can be found between the transactions/products.

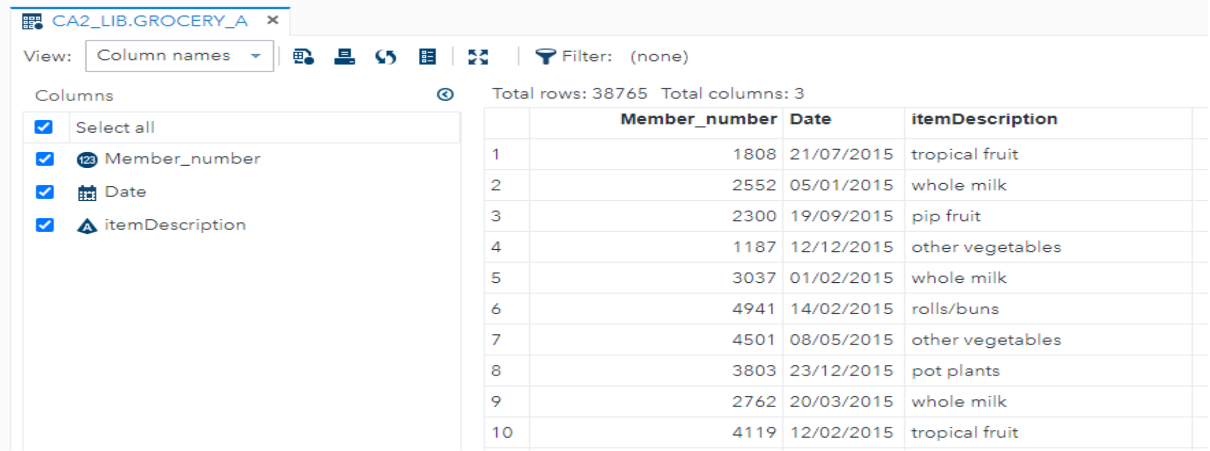
The benefits of such an analysis are to determine the strength of the buying patterns and how/if this should influence the layout and presentation of goods in our hypothetical retail grocery store.

Is there insight in this process that can help both the customer and retailer shop faster and smarter? If certain items, or groups of items, are frequently bought together then there is a solid logic in placing these items closer together physically. For an e-commerce experience, it might mean reducing the number of clicks to move between items that are highly related in terms of purchasing patterns.

2. Data Exploration & Descriptive Analytics

This Kaggle grocery transaction dataset is relatively straightforward in structure. It contains **3** columns and **38,765** rows.

After selecting this dataset, the first action was to load it into SAS Studio and take a look at how the CSV file was imported, and data types assigned.



Total rows: 38765 Total columns: 3

	Member_number	Date	itemDescription
1	1808	21/07/2015	tropical fruit
2	2552	05/01/2015	whole milk
3	2300	19/09/2015	pip fruit
4	1187	12/12/2015	other vegetables
5	3037	01/02/2015	whole milk
6	4941	14/02/2015	rolls/buns
7	4501	08/05/2015	other vegetables
8	3803	23/12/2015	pot plants
9	2762	20/03/2015	whole milk
10	4119	12/02/2015	tropical fruit

Fig 1 – Grocery Dataset Imported Into SAS Studio.

Each row represents a purchase of one item on a given day by a customer.

The layout of our dataset;

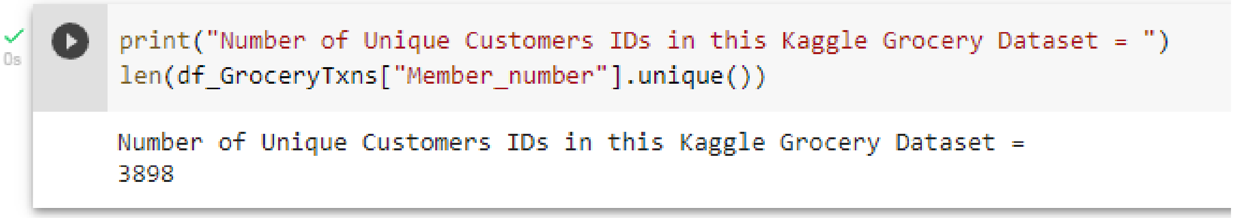
- ***Member_Number***: Unique number ID for customer.
- ***Date***: The purchase date of the transaction.
- ***ItemDescription***: Text detail of the single grocery item itself.

The 38,765 rows in the dataset each represent a unique transaction on one item, but many items are bought together by customers on a given day.

To better understand the information Kaggle provides an overview that in the dataset there are;

- **728** unique dates when grocery shopping took place.
- **167** unique items, which are sold by the shop.

A number of 3898 unique customers can be determined by a quick check on the unique number of customer IDs in the dataset. Below is a code snippet after the dataset was loaded into a Google Colab Python Notebook;

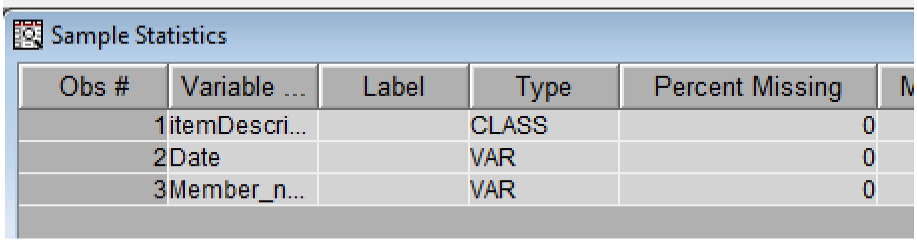


```
print("Number of Unique Customers IDs in this Kaggle Grocery Dataset = ")
len(df_GroceryTxns["Member_number"].unique())
```

```
Number of Unique Customers IDs in this Kaggle Grocery Dataset =
3898
```

Fig 2 – Number of Unique Customers in Grocery Dataset

After loading the data into SAS Enterprise Miner (EM), a quick check on the new data source shows that there are no missing values.



Sample Statistics

Obs #	Variable ...	Label	Type	Percent Missing	M
1	itemDescri...		CLASS	0	
2	Date		VAR	0	
3	Member_n...		VAR	0	

Fig 3 – Check on Imported Data for Grocery Data Source in SAS EM

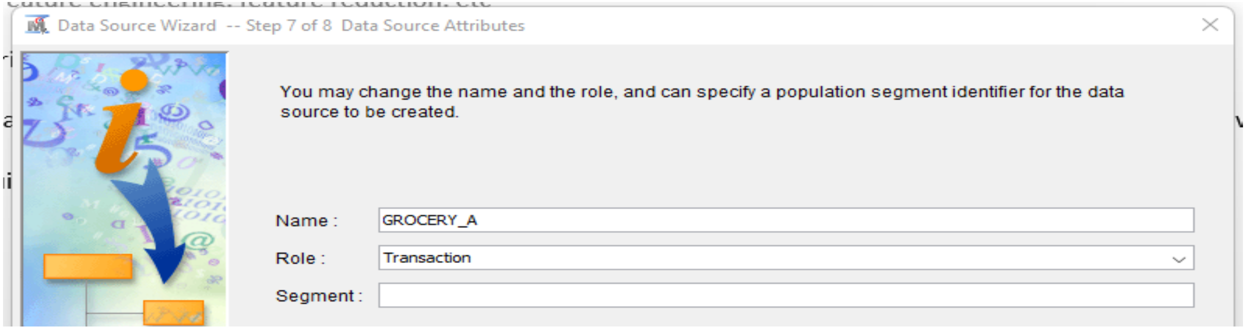
3. Data Preparation

Importing the Grocery ***csv*** file into SAS Studio converted the attributes into the correct data type.

It was then relatively straightforward to transfer the data as a Data Source into the project library in Enterprise Miner (EM) and, as Section 2 of this report section shows, the data was fully intact (no missing attributes).

To execute an Association Rules analysis some changes are required in SAS EM to the Data Source, and its attributes.

The Data Source 'role' was changed to 'Transaction';



Data Source Wizard -- Step 7 of 8 Data Source Attributes

You may change the name and the role, and can specify a population segment identifier for the data source to be created.

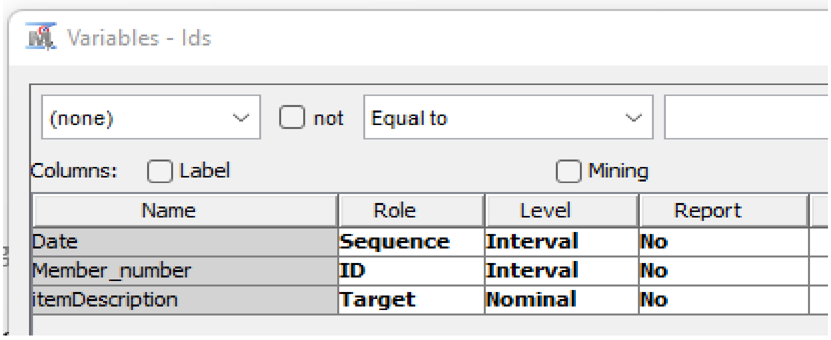
Name : GROCERY_A

Role : Transaction

Segment :

Fig 4 – Change Grocery Data Source Role to 'Transaction'

During the importing of the Data Source into the SAS EM project, the attributes in the Grocery dataset are also altered.



Variables - Ids

(none) □ not Equal to

Columns: □ Label □ Mining

Name	Role	Level	Report
Date	Sequence	Interval	No
Member_number	ID	Interval	No
itemDescription	Target	Nominal	No

Fig 5 – Update Column Metadata in Grocery Data Source

The column for 'itemDescription' contains the element for which we are trying to establish patterns with our Association Rules analysis, hence it is assigned the *Role* of 'Target'.

4. Details of Algorithms & Configurations

A diagram is created in SAS EM and the amended Data Source with the Grocery transaction information is added.

An 'Association' node is then connected to perform the actual association rules analysis.

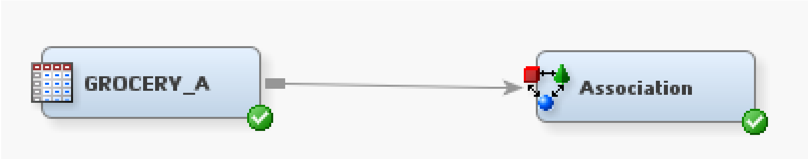
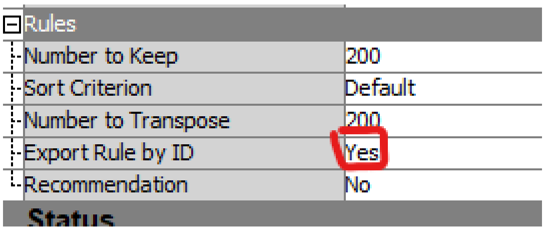


Fig 6 - Connecting Association Node to Data Source.

Most of the default settings in the node can be left unchanged, based primarily on guidance from DM lecture lab notes. The ‘Export Rule by ID’ is changed and set to ‘Yes’;

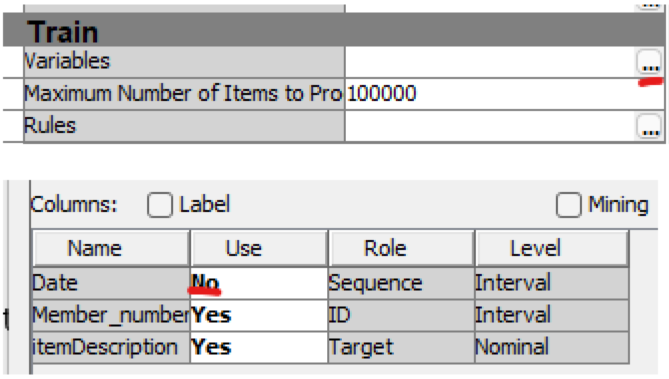


⊟Rules	
Number to Keep	200
Sort Criterion	Default
Number to Transpose	200
Export Rule by ID	Yes
Recommendation	No
Status	

Fig 7 – Change Export ID Rule

This allows the actual Association Rules to be saved. The Rules Description table will be made available in the ‘Results’ window display for review.

The last configuration change is to select the ‘Variables’ option under the ‘Train’ set of settings and update so that the ‘Date’ attribute is not used in the subsequent analysis.



Train	
Variables	...
Maximum Number of Items to Pro	100000
Rules	...

Columns: □ Label □ Mining

Name	Use	Role	Level
Date	**No**	Sequence	Interval
Member_number	**Yes**	ID	Interval
itemDescription	**Yes**	Target	Nominal

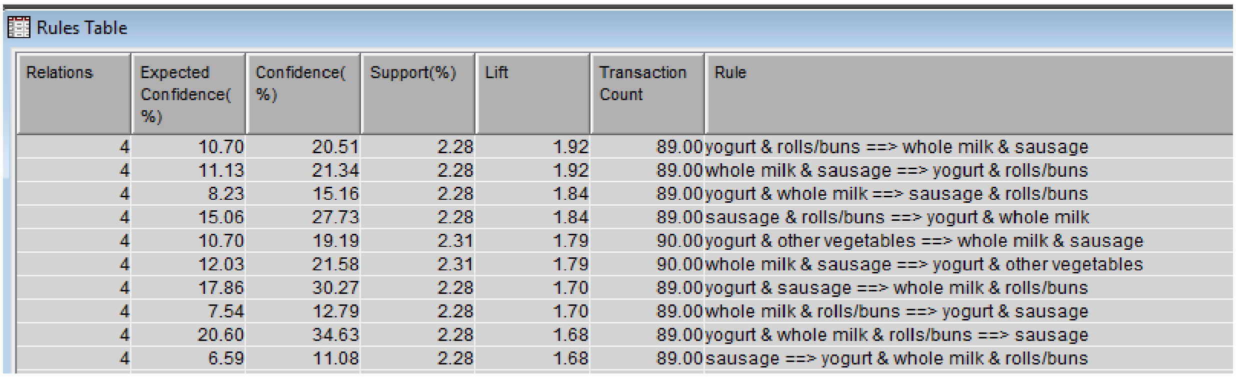
Fig 8 – Set ‘Date’ Attributes To ‘No’

5. Model Performance Metrics & Evaluation of Results

After the ‘Association Node’ is run we can assess the ‘Results’.

Rules Tables

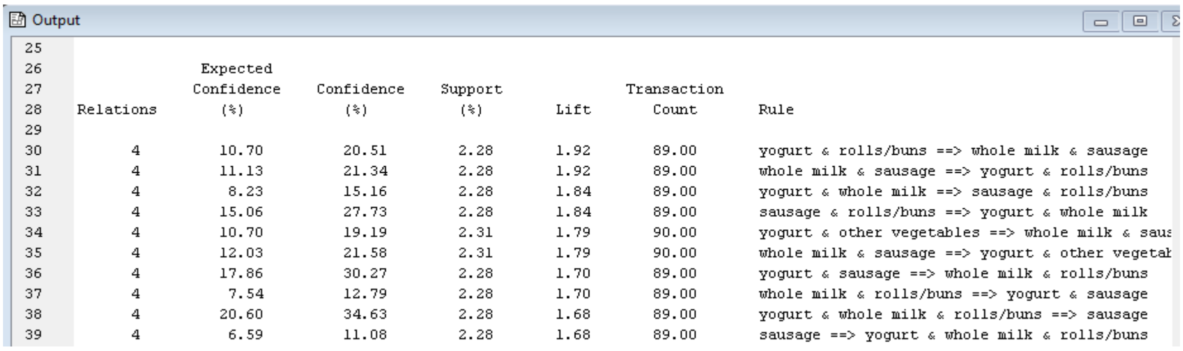
The first set of Result to assess are the ‘Rules Table’ and the ‘Output’, which is essentially the same data in different formats.



Rules Table

Relations	Expected Confidence(%)	Confidence(%)	Support(%)	Lift	Transaction Count	Rule
4	10.70	20.51	2.28	1.92	89.00	yogurt & rolls/buns ==> whole milk & sausage
4	11.13	21.34	2.28	1.92	89.00	whole milk & sausage ==> yogurt & rolls/buns
4	8.23	15.16	2.28	1.84	89.00	yogurt & whole milk ==> sausage & rolls/buns
4	15.06	27.73	2.28	1.84	89.00	sausage & rolls/buns ==> yogurt & whole milk
4	10.70	19.19	2.31	1.79	90.00	yogurt & other vegetables ==> whole milk & sausage
4	12.03	21.58	2.31	1.79	90.00	whole milk & sausage ==> yogurt & other vegetables
4	17.86	30.27	2.28	1.70	89.00	yogurt & sausage ==> whole milk & rolls/buns
4	7.54	12.79	2.28	1.70	89.00	whole milk & rolls/buns ==> yogurt & sausage
4	20.60	34.63	2.28	1.68	89.00	yogurt & whole milk & rolls/buns ==> sausage
4	6.59	11.08	2.28	1.68	89.00	sausage ==> yogurt & whole milk & rolls/buns

Fig 9 – Rule Table



Output

```

                 Expected
                Confidence    Confidence    Support            Transaction
Relations          (%)           (%)          (%)        Lift      Count        Rule


    4            10.70         20.51         2.28       1.92      89.00       yogurt & rolls/buns ==> whole milk & sausage
    4            11.13         21.34         2.28       1.92      89.00       whole milk & sausage ==> yogurt & rolls/buns
    4             8.23         15.16         2.28       1.84      89.00       yogurt & whole milk ==> sausage & rolls/buns
    4            15.06         27.73         2.28       1.84      89.00       sausage & rolls/buns ==> yogurt & whole milk
    4            10.70         19.19         2.31       1.79      90.00       yogurt & other vegetables ==> whole milk & saus
    4            12.03         21.58         2.31       1.79      90.00       whole milk & sausage ==> yogurt & other vegetab
    4            17.86         30.27         2.28       1.70      89.00       yogurt & sausage ==> whole milk & rolls/buns
    4             7.54         12.79         2.28       1.70      89.00       whole milk & rolls/buns ==> yogurt & sausage
    4            20.60         34.63         2.28       1.68      89.00       yogurt & whole milk & rolls/buns ==> sausage
    4             6.59         11.08         2.28       1.68      89.00       sausage ==> yogurt & whole milk & rolls/buns
```

Fig 10 – Results Output Table

Both of these diagrams are the Association Rules that shows the strongest links between grocery store items or groups of items. The rules are in descending order of 'Lift'. (Only the very highest entries are displayed here).

The '*Support*' value shows the popularity of a grocery item (or a group of items) in terms of the numbers of customer transaction that selected that item(s).

'*Confidence*' represents the likelihood of the RHS (Y) item being purchased if the LHS (x) has already been bought. Thus, from the above table, there is a 20.5% chance of *whole milk and sausage* being bought if the customer has already purchased *yogurt and rolls/buns.* The problem with such a metric is that very popular items can distort the confidence value and suggest a correlation that is stronger than it actually is.

'Lift' attempts to counteract the problems with using the 'Confidence' measure by adjusting the probability to counter for very popular RHS (Y) items. It can thus be a more accurate reflection of the correlation between items, or groups of items, and this is why the above tables show a scale based on 'Lift' values.

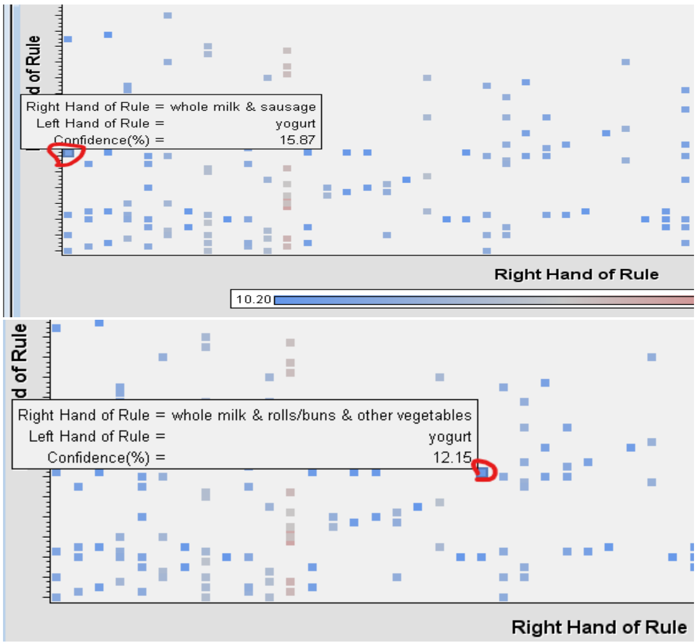
If the Lift value is above 1 then there is an increasing association. The closer to 1 the lower the association, and values lower than 1 indicate a negative correlation.

The '*yogurt & rolls/buns ==> whole milk & sausage*' is the strongest association in the dataset and is followed in the diagrams above by its reciprocal rule.

LHS v RHS

The LHS – RHS Rule Matrix contains a great deal of data but is colour coded to help with ad-hoc analysis.

Taking a quick example of Yogurt as a LHS Value;



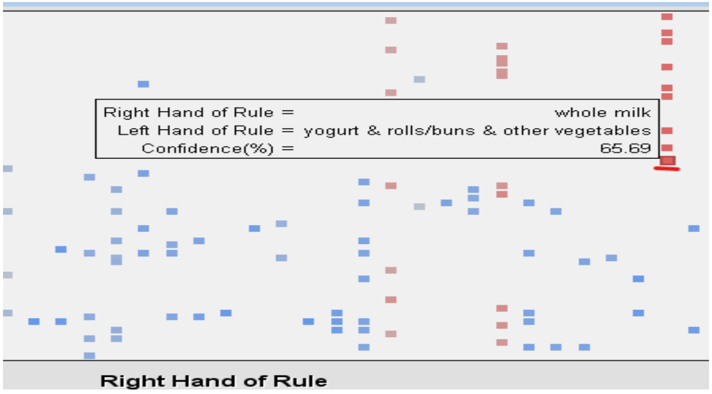
Right Hand of Rule = whole milk & sausage
Left Hand of Rule = yogurt
Confidence(%) = 15.87

Right Hand of Rule = whole milk & rolls/buns & other vegetables
Left Hand of Rule = yogurt
Confidence(%) = 12.15

Fig 11 - Yogurt Confidence

As one moves along a straight line from left to right, the changing degrees of confidence that RHS items have in relation to Yogurt (LHS) can be shown. All these cells are pale blue indicating a lower degree of confidence.

Looking at Whole Milk, as one moves from bottom to top the confidence values for Whole Milk to other items can be seen. These are darker red dots and of much more interest to the store owner in terms of strong product correlation.



Right Hand of Rule = whole milk
Left Hand of Rule = yogurt & rolls/buns & other vegetables
Confidence(%) = 65.69

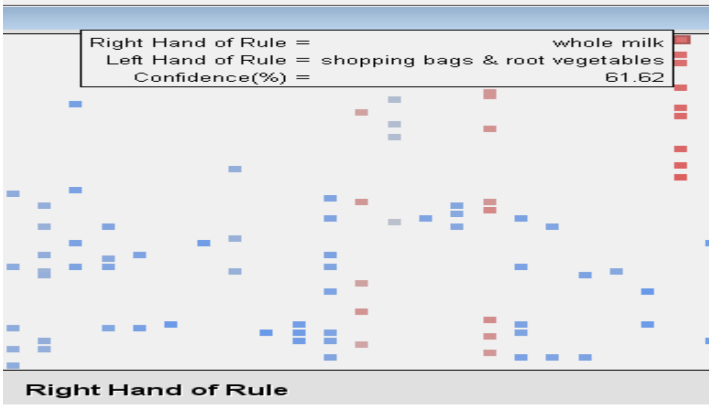

Right Hand of Rule = whole milk
Left Hand of Rule = shopping bags & root vegetables
Confidence(%) = 61.62
Right Hand of Rule

Fig 12 – Whole Milk Confidence

One further output of interest from SAS EM is a Link Graph to visually show the strength of relationships between the store items. This next diagram is somewhat hard to read but emphasises the analysis from the other SAS EM Result outputs.

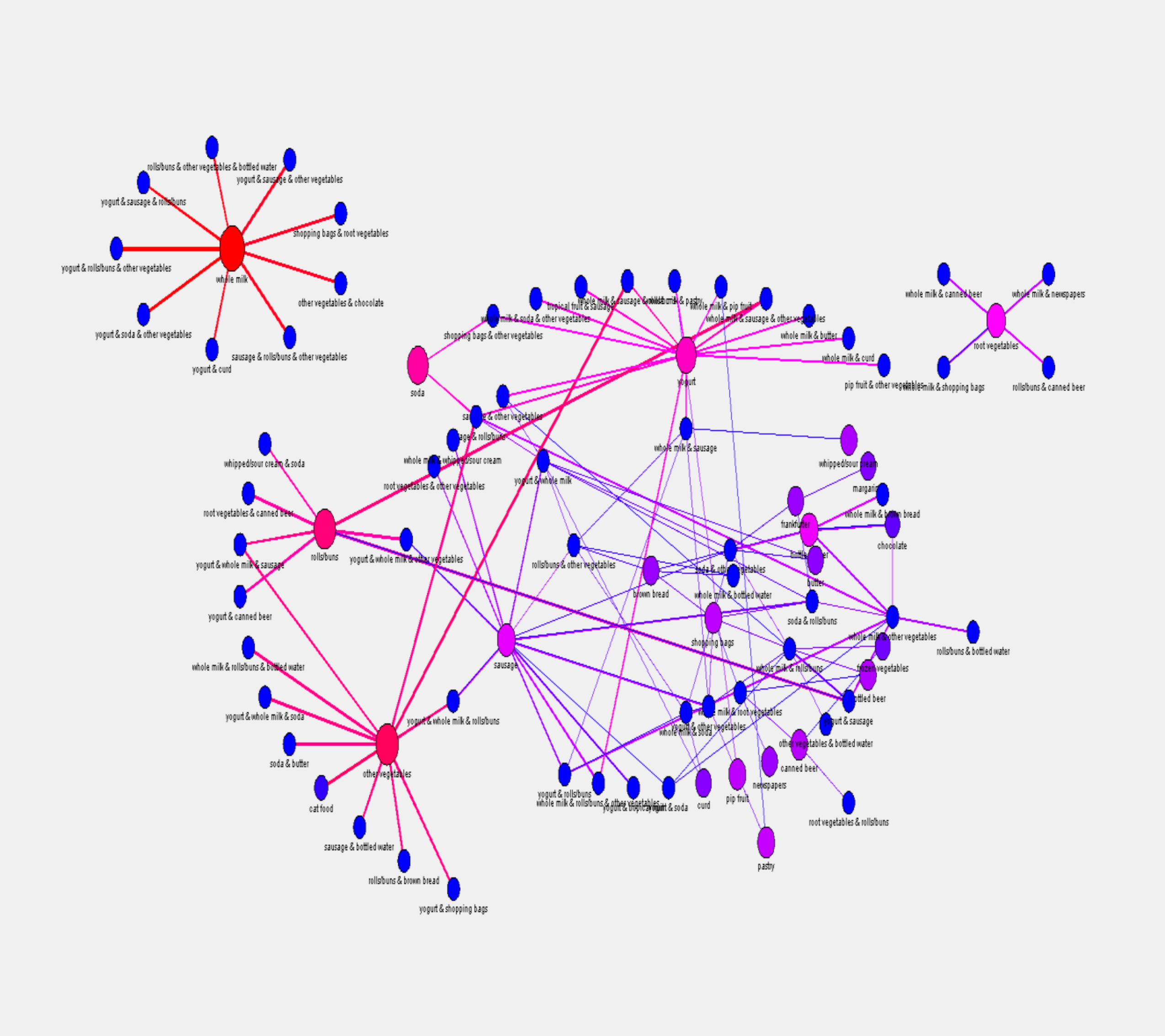

rolls/buns & other vegetables & bottled water
yogurt & sausage & other vegetables
yogurt & sausage & rolls/buns
shopping bags & root vegetables
yogurt & rolls/buns & other vegetables
whole milk
other vegetables & chocolate
yogurt & soda & other vegetables
sausage & rolls/buns & other vegetables
yogurt & curd
tropical fruit & sausage
whole milk & pip fruit
whole milk & soda & other vegetables
whole milk & sausage & other vegetables
shopping bags & other vegetables
whole milk & butter
whole milk & curd
soda
yogurt
pip fruit & other vegetables
whole milk & canned beer
whole milk & newspapers
root vegetables
whole milk & shopping bags
rolls/buns & canned beer
whole milk & whipped/sour cream
root vegetables & other vegetables
yogurt & whole milk
whole milk & sausage
whipped/sour cream
whipped/sour cream & soda
root vegetables & canned beer
yogurt & whole milk & sausage
rolls/buns
yogurt & whole milk & other vegetables
yogurt & canned beer
rolls/buns & other vegetables
brown bread
frankfurter
whole milk & brown bread
chocolate
butter
soda & other vegetables
whole milk & bottled water
soda & rolls/buns
shopping bags
sausage
whole milk & other vegetables
rolls/buns & bottled water
whole milk & rolls/buns & bottled water
whole milk & rolls/buns
frozen vegetables
yogurt & whole milk & soda
yogurt & whole milk & rolls/buns
whole milk & root vegetables
yogurt & other vegetables
whole milk & soda
yogurt & sausage
soda & butter
other vegetables
other vegetables & bottled water
cat food
canned beer
newspapers
yogurt & rolls/buns
whole milk & rolls/buns & other vegetables
yogurt & soda
curd
pip fruit
root vegetables & rolls/buns
sausage & bottled water
rolls/buns & brown bread
pastry
yogurt & shopping bags

Fig 13 - Link Graph output from SAS EM

Overall Observations from Association Rules Analysis on Grocery Dataset

- *Whole Milk* is the item with the strongest associations with other items. It is represented by the deepest red dot in the Link Graph and tops the Rules Tables. Not surprisingly it associates strongly with items such as vegetables and yogurt.

- Various vegetable types, yogurt, rolls/buns, and soda all feature with strong relationships to other items.

- Soda is perhaps the least 'healthy' of the main items, but the strength of the associations might warrant a review of its location on shelves within the physical retail outlet.

Adding a Chi Square Statistic

A SAS EM research paper in 2012[1] proposed adding a SAS Code Utility node with Chi Squared analysis code to a Grocery Market Basket analysis. The purpose was to provide an additional Results output to show if any of the association rules were not statistically significant, and hence if the overall set of rules could be reduced to a smaller size.

The source code was taken from the appendix of that research paper and added to our Associations Rules diagram in SAS EM.

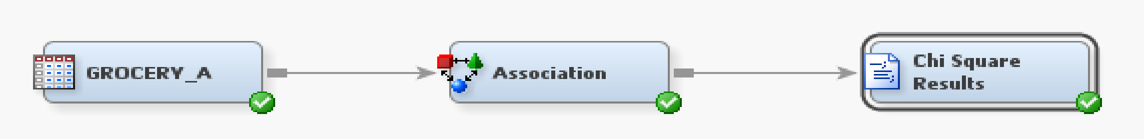
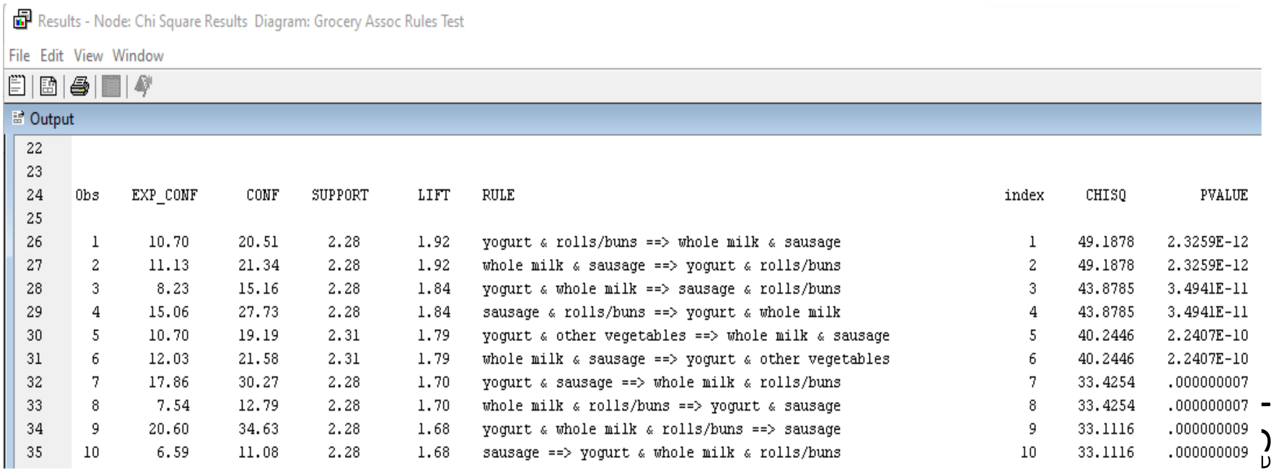


Fig 14 - Adding SAS Code Node (which has been renamed) to Diagram.

Although this was an interesting exercise, there were no rules in the 200-rule set in this experiment with a high enough *PValue* to declare them as not statistically significant (at alpha = .05).


Results - Node: Chi Square Results Diagram: Grocery Assoc Rules Test
File Edit View Window
Output

```
Obs  EXP_CONF    CONF   SUPPORT    LIFT   RULE                                             index    CHISQ       PVALUE

  1    10.70    20.51     2.28     1.92   yogurt & rolls/buns ==> whole milk & sausage         1   49.1878   2.3259E-12
  2    11.13    21.34     2.28     1.92   whole milk & sausage ==> yogurt & rolls/buns         2   49.1878   2.3259E-12
  3     8.23    15.16     2.28     1.84   yogurt & whole milk ==> sausage & rolls/buns         3   43.8785   3.4941E-11
  4    15.06    27.73     2.28     1.84   sausage & rolls/buns ==> yogurt & whole milk         4   43.8785   3.4941E-11
  5    10.70    19.19     2.31     1.79   yogurt & other vegetables ==> whole milk & sausage   5   40.2446   2.2407E-10
  6    12.03    21.58     2.31     1.79   whole milk & sausage ==> yogurt & other vegetables   6   40.2446   2.2407E-10
  7    17.86    30.27     2.28     1.70   yogurt & sausage ==> whole milk & rolls/buns         7   33.4254   .000000007
  8     7.54    12.79     2.28     1.70   whole milk & rolls/buns ==> yogurt & sausage         8   33.4254   .000000007
  9    20.60    34.63     2.28     1.68   yogurt & whole milk & rolls/buns ==> sausage         9   33.1116   .000000009
 10     6.59    11.08     2.28     1.68   sausage ==> yogurt & whole milk & rolls/buns        10   33.1116   .000000009
```

Fig 15 – Output From Chi Squared Analysis (excerpt of top ten rules)

6. Comparison with Other Research

Enhance the decision-making process around business rules

A paper published by Faron and Chakraborty at the SAS Global Forum in 2012[1] suggested a way to add the Pearson's Chi Squared statistic to the output from the SAS EM *Association* node.

The addition of this test would enhance the quality of the data analysis by showing which rules are statistically significant, therefore adding an additional metric the evaluate the importance of given association rules.

I implemented this Chi Squared analysis, but it did not yield any obvious benefits (or none which I could determine).

Representing Rules: Visual Clutter

The Link Graph output in SAS EM is a useful at a glance tool to visually represent the associations rules in our grocery database. However, research on similar grocery datasets, such as that by Hahsler, M., Hornik, K. and Reutterer, T., 2016[2], highlights the challenge that these graphs easily become cluttered as the number of rules grow. (This would be true of any rule set, not just the typical grocery basket analysis being conducted in this section of the assignment).

If a visual representation of the association rules is an important outcome, then there are tools such as ***arulesViz*** that offer more sophisticated interactive functionality. For the purposes of this assignment the SAS EM graph tools were deemed sufficient.

What to do with these Association Rules?

This dataset is a relatively straightforward representation of grocery purchases, and the common benefit is often considered to be a more effective physical retail shop layout.

A key point made in a 2018 Towards Data Science article[3] is that Association Rules look at lists of items with unique transaction ID from many users, and studies these lists as a block. This is not an approach that generates a recommendation for one *specific* user. That said, research I found on the role of Association Data Mining and E-Commerce website structure[4] shows how the selection of *antecedent* lists can be used to meaningfully direct a single user to different web pages, where they are most likely to find the products for which they are looking.

A follow-on challenge for me would be to repeat this exercise with a similar but extended dataset that looked at product groupings in the *itemsets*, and also considered user profiles.

7. References

[1] Faron, M. and Chakraborty, G., 2012. *Easily Add Significance Testing to your Market Basket Analysis in SAS® Enterprise Miner*. [online] Citeseerx.ist.psu.edu. Available at: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.421.1785&rep=rep1&type=pdf> [Accessed 5 January 2022].

[2] Hahsler, M., Hornik, K. and Reutterer, T., 2016. Implications of Probabilistic Data Modelling for Mining Association Rules. *From Data and Information Analysis to Knowledge Engineering*, [online] pp.598-605. Available at: <https://link.springer.com/content/pdf/10.1007/s11573-016-0822-8.pdf> [Accessed 5 January 2022].

[3] Garg, A., 2018. *Complete guide to Association Rules (1/2)*. [online] TowardsDataScience. Available at: <https://towardsdatascience.com/association-rules-2-aa9a77241654> [Accessed 31 December 2021].

[4] Omari, A., Conrad, S. and Alcic, S., 2007. Designing a Well-Structured E-Shop Using Association Rule Mining. *2007 Innovations in Information Technologies (IIT)*, [online] Available at: <https://ieeexplore.ieee.org/abstract/document/4430429> [Accessed 5 January 2022].

TASK 4 – *Ethics and the user of Data Science/ML/AI*

Task 4-1 : Stop The Killer Robots – Autonomous Drone Warfare

1. Overview of problem

This title may sound like a bad 'B-Movie' but the *Campaign to Stop Killer Robots* (https://www.stopkillerrobots.org/) is an umbrella framework for 180+ organisations that is concerned about the growing potential threat of autonomous weapons systems.

In this section of the assignment, I am considering the ethical and legal issues raised by groups like *Stop The Killer Robots* (STKR) that are associated with this extreme end of the spectrum when it comes to autonomy in technology.

What can, and should, be done to ensure that the artificial intelligence development and processes underpinning 'killer robots' is accountable and free from abuse?

The question becomes less and less academic each day. We already have present day examples of quasi-autonomous weapons in the field, such as Israel's *Harpy* anti-radar 'fire-and-forget' drone. These, and ongoing military AI development across the globe, raise an ongoing moral dilemma around such technology[1].

2. Ethical and Legal Challenges

Paul Scharre, a former US-Army Ranger, and a director at the New American Security 'think-tank', wrote in 2019 in his book: *Army of None: Autonomous Weapons and the Future of War* that the Pentagon needed to shift its thinking on artificial intelligence[2].

Scharre distilled his concerns down to two kinds of legal and ethical questions;

1. Machine permissibility. What is the system allowed to do on its own?

2. Machine accountability. Who takes initial (and ultimate) responsibility for what the system does on its own?

In February 2020, as a response to these types of concerns, the US DoD rolled out a list of five AI ethical principles to govern its work in this area[3]:

1. Responsibility and good judgement applied by military personnel in the use of AI capabilities.
2. Equitable. Bias is minimized.
3. Traceable. Reasons for AI decisions can be understood.
4. Reliable. Systems tested, secured, and robust.
5. Governable. The ability exists to easily disengage in the case of unintended behaviour.

However, the concern from the *Campaign to Stop Killer Robots* is that these effectively remain guidelines and could ultimately be used by human actors to avoid taking legal (or moral) responsibility for the actions of autonomous weapons systems[4].

The campaign highlights that a new proposed international treaty to prohibit and restrict 'killer robots' has been endorsed by dozens of countries[5]. Despite this, the major powers remain resistant to new treaties, preferring to look at existing legislation and regulation [6][7]. This fuels scepticism in many that the desire to be first in the 'AI Arms Race' will lead to compromises in ethical standards.

3. Challenges for Data Scientist

The US DoD have declared that they want to integrate ethics into all aspects of their AI test and evaluation processes[8], and thus have outlined policies for their data engineers.

In the need to be equitable and traceable, the AI test harnesses must be able to identify algorithmic bias. It must be clear what data elements are contributing to a systems decision. If a system is literally going to be targeting an individual, or group of individuals, it must be clear what criteria the machine learning model is using to make that decision.

Even for the DoD a major challenge is that AI testing is still heavily dependent on manual assessment. There is widespread engagement with the private sector and academia, but this is still seen as an area of concern. There is a lack of sophisticated toolkits to test AI-driven systems, in the view of the US DoD.

The disengagement mechanism appears to be more of a general engineering challenge in terms of capability, rather than one unique to data engineers. 'Pulling the plug' quickly and effectively, if needed, in the event of a suspicious decision requires a generally well-built system architecture.

Other US government departments are impressed with the AI techniques being deployed by the DoD and seek to emulate them in their own ethical artificial intelligence strategies[9].

4. Reflections

Many voices in the *Campaign to Stop The Killer Robots* advocate an outright ban on AI technologies being used to create autonomous weapons systems. Professor Noel Sharkey has passionately argued that computers should never be in the business of killing people[4].

However, Is an outright ban even remotely practical? Many in the military today believe that such a ban is impossible [10].

Crucially, there are other voices in the STKR organisation, with both academic and military backgrounds, that push for governments to, at the very least, adopt the US DoD ethical principles and then enshrine this process in multi-lateral treaty agreements. Critically, it should be paramount that 'permanent significant human control' always remains in place [11].

To me, it seems that, just like with Nuclear and Chemical weapons before them, AI-based weapons need to be comprehensively covered under dedicated international arms treaties. Such weapons will proliferate and become a great deal harder to count and verify, but pressure needs to be brought to bear on the major powers to recognise the genuine concerns of humanity in the face of 'killer robots'.

5. References

[1] Winter, C., 2017. *'Killer robots': autonomous weapons pose moral dilemma | DW | 14.11.2017*. [online] DW.COM. Available at: <https://p.dw.com/p/2nT6O> [Accessed 11 December 2021].

[2] Scharre, P., 2019. *Army of None: Autonomous Weapons and the Future of War*. 1st ed. New York: W. W. Norton & Company.

[3] Barnett, J., 2020. *Pentagon adopts ethical principles for artificial intelligence*. [online] FedScoop. Available at: <https://www.fedscoop.com/dod-ai-ethics-principles/> [Accessed 11 December 2021].

[4] Wareham, M., 2020. *Robots Aren't Better Soldiers than Humans*. [online] Hrw.org. Available at: <https://www.hrw.org/node/376854/printable/print> [Accessed 11 December 2021].

[5] Wareham, M., 2020. *Killer Robots: Growing Support for a Ban*. [online] Human Rights Watch. Available at: <https://www.hrw.org/news/2020/08/10/killer-robots-growing-support-ban> [Accessed 11 December 2021].

[6] Klane, M., 2018. "U.S., Russia Impede Steps to Ban 'Killer Robots.'". *Arms Control Today*, [online] 48(8), pp.31-33. Available at: <https://www.jstor.org/stable/90025262> [Accessed 11 December 2021].

[7] Bowcott, O., 2015. *UK opposes international ban on developing 'killer robots'*. [online] The Guardian. Available at: <https://www.theguardian.com/politics/2015/apr/13/uk-opposes-international-ban-on-developing-killer-robots> [Accessed 11 December 2021].

[8] Barnett, J., 2020. *How the DOD is developing its AI ethics guidance*. [online] FedScoop. Available at: <https://www.fedscoop.com/jaic-alka-patel-ai-ethics/> [Accessed 11 December 2021].

[9] Nyczepir, D., 2020. *HHS AI chief sees promise in emulating the DOD*. [online] FedScoop. Available at: <https://www.fedscoop.com/hhs-ai-office-jaic/> [Accessed 11 December 2021].

[10] Nast, C., 2020. *There's No Turning Back on AI in the Military*. [online] Wired. Available at: <https://www.wired.com/story/opinion-theres-no-turning-back-on-ai-in-the-military/> [Accessed 11 December 2021].

[11] Micha, L. and Farias, P., 2021. *The evolution of disruptive technologies and lethal autonomous weapons systems: considerations from the military field*. [online] Stopkillerrobots.org. Available at: <https://www.stopkillerrobots.org/wp-content/uploads/2021/09/The-evolution-of-disruptive-technologies-and.pdf> [Accessed 10 December 2021].

Task 4-2 : YouTube in 2021: Trying to Tame the Recommendation Engines That Radicalised Millions (By Accident).

1. Overview of problem

Guillaume Chaslot completed is doctorate in Machine Learning in 2010 and was then offered his dream job working at Google.

He began, along with other data scientists, working on the artificial intelligence algorithms that drove the YouTube recommender sidebar, and was excited by the possibilities[1].

By 2013, Guillaume had been fired from YouTube. He had advocated against the changes in the recommender engine that were inadvertently pushing mis-leading and hateful content at people, all with the intention of keep viewers hooked on YouTube.

The role of YouTube, and other social media platforms, in generating a radicalised subculture of viewers is well documented. The problem that YouTube grapples with in 2021 is how address the criticisms of policies in 2014-2018 and to be a platform that promotes diversity and truthfulness. Can it handle the test of weeding out the undesirable content, and has the challenge of the Covid workplace made it more difficult to meet this ethical objective?

2. Ethical and Legal Challenges

The fourth episode of the 2021 New York Times ‘Rabbit Hole’ podcast contains an interview with Susan Wojcicki, the CEO of YouTube. It focuses primarily on decisions driven by her nearly a decade ago to change the way the YouTube recommendation engine worked, and partial acknowledgement that this led to unexpected (and presumably undesirable) alt-right radicalisation of significant numbers of viewers[2].

The accusation laid against YouTube, both by external observers and former staff, is that around 2014 the company deliberately chose to refine its recommender algorithm with the express intention of increasing *watch time*.

A sophisticated neural network model would recommend, and actively promote in the side bar, a greater range and diversity of new videos based on past viewing history. Most critically, YouTube would continue to filter out content based on the obvious, and established, criteria for banning videos - violence, nudity, and profanity - but would do little to assess content beyond those measures. Thus, an increasing array of alt-right videos were being pushed out to viewers who has started searching for videos on topics such as history and self-help[1].

The alt-right videos, with their sensationalist titles, tended to generate more views and led certain viewers down a ‘rabbit hole’ of conspiracy theories and hate speech.

YouTube defended it actions at the time by declaring that it did deliberately did not take a partisan side in politics, and that it respected free speech.

However, the ethical issue continued to be that YouTube was providing a platform to hate speech and misinformation. More worryingly, the recommender engine was too often creating 'filter bubbles'[1]. The model was clever enough not to fixate on cat videos, if that was the initial user search topic. However, if the user looked at an Alex Jones diatribe the YouTube model would never try and counterbalance with something from Jon Stewart.

YouTube was feeding a serious imbalance in news media, and it took until late 2019 before it began to effectively acknowledge responsibility, and it became evident and visible that YouTube was getting serious about its ethical media stance.

Of course, as welcome as these changes are, it is worth remembering that free speech is desirable, and content should be available with opposing views. Is there a danger that YouTube might inadvertently suppress content just because it is unpopular?

3. Challenges for Data Scientist

Probably the most obvious challenge is volume. The YouTube Community Guideline site states that 100s of hours of content are uploaded every minute[3].

Compounding this metric is the fact that YouTube is still heavily reliant on human moderators. YouTube sent its workforce, including the 10,000+ moderators, home at the start of the Covid-19 pandemic and ramped up the scope and operations of the automated AI moderation system. However, these systems reported a significantly higher volume of 'false positives', taking down videos that were actually not in breach of guidelines and for which half of the decisions were subsequently reversed[4]. The dependency on human moderation became even more apparent when YouTube made the decision to re-introduce greater levels of human involvement in late 2020[5].

This highlights that the machines are not quite ready yet to replace the human element in content moderation, and that the assessment of videos in bulk remains a complex task in need of further solutions.

That said, the company states that, as of April 2021, 94% of content breaking its rules is caught by its AI systems, and most of those videos are removed before they have 10 views[6]. Clearly the situation is improving but difficulties remain.

4. Reflections

Susan Wojcicki's interview in 2020 with the New York Times[2] seemed to imply that, at the time, she doubted that changes to YouTube could do much harm. Politics seemed a very niche element of YouTube, with very low viewership.

However, the AI models that were built and deployed in 2012-2015 were designed to capture and keep attention as a primary objective. They worked extremely well and are seem by many as a contributor to the increased polarization and coarseness of global political debate in 2021.

I have two reflections;

The first is that work is still needed to enhance the AI models to prevent misinformation and hate speech being uploaded to platforms like YouTube. In December 2021, Meta announced innovations it was working on in the field of Few-Shot Learner (FSL), an AI technology to allow more rapid action on capturing harmful content[7]. Presumably, these are the types of approaches that YouTube will also embrace.

The second is related to the fact that since 2020 YouTube has been effectively saying that it will follow the establishment line on topics such as vaccines and hate crime. This is a positive move, but the 'establishment' is not always right. A recent example is the self-censorship of YouTube in Russia[8]. Is there not an argument that we should encourage YouTube, and other such platforms, to allow the occasional 'edgy' content?

5. References

[1] Roose, K. and Mills, A., 2021. *One: Wonderland (Published 2020)*. [online] Nytimes.com. Available at: <https://www.nytimes.com/2020/04/16/podcasts/rabbit-hole-internet-youtube-virus.html?> [Accessed 12 December 2021].

[2] Roose, K. and Miils, A., 2021. *Four: Headquarters (Published 2020)*. [online] Nytimes.com. Available at: <https://www.nytimes.com/2020/05/07/podcasts/rabbit-hole-youtube-susan-wojcicki-virus.html?showTranscript=1> [Accessed 12 December 2021].

[3] YouTube Community Guidelines and policies - How YouTube Works. 2021. *YouTube Community Guidelines and policies - How YouTube Works*. [online] Available at: <https://www.youtube.com/howyoutubeworks/policies/community-guidelines/#detecting-violations> [Accessed 12 December 2021].

[4] Lapowsky, I., 2020. *After sending content moderators home, YouTube doubled its video removals*. [online] Protocol — The people, power, and politics of tech. Available at: <https://www.protocol.com/youtube-content-moderation-covid-19#toggle-gdpr> [Accessed 12 December 2021].

[5] Kraus, R., 2020. *YouTube puts human content moderators back to work*. [online] Mashable. Available at: <https://mashable.com/article/youtube-human-content-moderation> [Accessed 12 December 2021].

[6] Heilweil, R., 2021. *YouTube says it's better at removing videos that violate its rules, but those rules are in flux*. [online] Vox. Available at: <https://www.vox.com/recode/2021/4/6/22368809/youtube-violative-view-rate-content-moderation-guidelines-spam-hate-speech> [Accessed 12 December 2021].

[7] Meta. 2021. *Meta's New AI System to Help Tackle Harmful Content | Meta*. [online] Available at: <https://about.fb.com/news/2021/12/metas-new-ai-system-tackles-harmful-content/> [Accessed 12 December 2021].

[8] Bergin, M. and Chang, E., 2021. *YouTube CEO Says Google Sees Free Speech as Core Value in Russia*. [online] Bloomberg.com. Available at: <https://www.bloomberg.com/news/articles/2021-09-26/youtube-ceo-says-google-sees-free-speech-as-core-value-in-russia> [Accessed 12 December 2021].