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 |  |
| Date |  |

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](https://www.brewersfriend.com/search/) allows homebrew enthusiasts to upload and share their own recipes. A Kaggle project is located here: [Brewer's Friend Beer Recipes | Kaggle](https://www.kaggle.com/jtrofe/beer-recipes), 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 ‘Custer’ 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.

1. **Data Exploration & Descriptive Analytics**

Include any data insights discovered

*Basic 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:

Graphical user interface, text, application

Description automatically generated

Fig <n> Python – Dataset Dimensions and Colum List

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

Table

Description automatically generated

Fig <n> 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 build 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.

Graphical user interface, application

Description automatically generated

Fig <n> SAS EM – Filter Only on America IPA

I chose to do this 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. In general, the numerical data looks to be well set for accurate clustering later in this task.

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 <n> 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 attributes upon which our cluster analysis should be built.

Diagram

Description automatically generated

Fig <n> Homebrew Process

* *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**[1]**.

*Brief Analysis of Key Clustering Numerical Attributes*

The filtering of the homebrew data to only American IPA has removed several 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.

1. **Data Preparation**

Include details of any data cleaning, transformations, data enrichment, feature engineering, feature reduction, etc

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.

<Screen shot of 2nd Graph Explore in EM>

Fig <n> 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 confirm with data range values that an experienced homebrew would expect to see in each recipe**[2]**.

1. **Details of Algorithms & Configurations**

*Additional Preparation for Cluster Analysis*

Setting up SAS EM for Clustering analysis is relatively straightforward.

<Diagram of Cluster/Segment Profile>

Fig <n> 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:

**Table

Description automatically generated**

Fig <n> SAS EM Cluster Node Setting for Standardisation

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

**Graphical user interface, text, application

Description automatically generated**

Fig <n> Python Code for Standardisation

How Many Clusters?

Running the SAS EM ‘Cluster’ node with the automation setting for numbers of clusters generates nearly **50** clusters in the node results.

<Diagram of Cluster Setting for Cluster Numbers>

Fig <n> SAS EM Cluster Node Setting for Standardisation

SAS EM has found a pattern to group the American IPA recipe data into 50 groups based on common characteristics. In practice. this is an unwieldy number with which to work and process to 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.

Graphical user interface, text, application, email

Description automatically generated

Fig <n> Python Code for KMeans

The following graph is generated

**Chart, line chart

Description automatically generated**

Fig <n> The Elbow Method Graph

The second ‘bend’ in the elbow represents the ideal number of clusters. In this case the optimal cluster number appears to be **7**.

1. **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.

<Diagram of Cluster Setting for Standardisation>

Fig <n> 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.

<Diagram of Cluster Setting for Standardisation>

Fig <n> 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 mode to gain a better understanding of how the clusters have been created.

<Diagram of Segment Profile Node>

Fig <n> SAS EM Segment Profile Node

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.

Chart, histogram

Description automatically generated

Fig <n> 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.

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

Chart, bar chart

Description automatically generated

Fig <n> 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 is extended with additional filter nodes to isolate cluster 4 entries, and then remove any rows that have missing categorical attributes describing the brewing process.

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

Graphical user interface

Description automatically generated with low confidence

Fig <n> Extracting Clustered Data into EXCEL

1. **Comparison with other Research & Reflections**

1 – Feature Engineering with Domain Knowledge

The Kaggle Notebook by Dereck Bearsong**[2]**, 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 knowledge repositories 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 on AI-techniques to develop machine learning-built IPA recipe templates**[3]**. The source data for this work was also a 70K+ dataset of publicly available recipes, which then focused on IPA beers. The concepts were challenging, 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 drew from the Brewers Friend dataset and conducted an analysis of the impact of regional ingredients on beers**[4]**. 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. In fact, the dataset has only 170 and 229 genuinely unique malts and hops respectively. This has 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 values (as in the UCD study). It would be interesting to see if this impacted on the outcome from the Clustering analysis.

1. **References**

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

[2] 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].

[3] 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].

[4] 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 - *<insert select k Name here e.g. Association Rules Problem>*

1. **Definition of Problem**

Clearly state the problem definition, what type of data mining task is it, where was the data set sourced from, etc.

1. **Data Exploration & Descriptive Analytics**

Include any data insights discovered

1. **Data Preparation**

Include details of any data cleaning, transformations, data enrichment, feature engineering, feature reduction, etc

1. **Details of Algorithms & Configurations**
2. **Model Performance Metrics & Evaluation of Results**
3. **Comparison with other Research**

Compare your results to at least three other researchers (maximum of five) who used the same data set. What lessons did you learning from doing this? How can your work be improved? Did you include any improvements in your work and what impact did it have?

1. **References**

Use the IEEE Referencing style. See this guide for details. <https://libraryguides.vu.edu.au/ieeereferencing/gettingstarted>

# TASK 3 - *<insert select Task Name here e.g. Time Series Analysis Problem>*

1. **Definition of Problem**

Clearly state the problem definition, what type of data mining task is it, where was the data set sourced from, etc.

1. **Data Exploration & Descriptive Analytics**

Include any data insights discovered

1. **Data Preparation**

Include details of any data cleaning, transformations, data enrichment, feature engineering, feature reduction, etc

1. **Details of Algorithms & Configurations**
2. **Model Performance Metrics & Evaluation of Results**
3. **Comparison with other Research**

Compare your results to at least three other researchers (maximum of five) who used the same data set. What lessons did you learning from doing this? How can your work be improved? Did you include any improvements in your work and what impact did it have?

1. **References**

Use the IEEE Referencing style. See this guide for details. <https://libraryguides.vu.edu.au/ieeereferencing/gettingstarted>

# TASK 4 - *<insert select Task Name here e.g. Data Ethical Issues >*

## Task 4-1 : <Title of Case Study)

1. **Overview of problem**
2. **Ethical and Legal Challenges**
3. **Challenges for Data Scientist**
4. **Reflections**
5. **References**

Use one of the commonly used References and Citation formats.

## Task 4-1 : <Title of Case Study)

1. **Overview of problem**
2. **Ethical and Legal Challenges**
3. **Challenges for Data Scientist**
4. **Reflections**
5. **References**

Use one of the commonly used References and Citation formats.