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 recipe 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 scrapped 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 the 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 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> SAS EM – EXPLORE View of Dataset Attributes

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

*Focus on American IPA First*

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.

However, 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.

The Kaggle reference data indicates that both the ‘StyleID ‘ and ‘Style’ attributes can be used to filter on beer type. American IPA has a ‘StyleID; number of ‘7’. Hence, we introduce a filer node in SAS EM (and Python code) to reduce our working dataset to just American IPA recipes.

Graphical user interface, application

Description automatically generated

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

I chose to do this 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 rows, but it can be seen in Fig <n> that the most common value for most categorical attributes is ‘N/A’. We will return to this in the final stages of this task.

*What Attributes are Important for this Clustering Task?*

Reference to: [Quick visualization & analysis of Homebrew Recipes | Kaggle](https://www.kaggle.com/blasterbrewmaster/quick-visualization-analysis-of-homebrew-recipes)

Reference to picture source…

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

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 & Reflections**

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 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.