School of Science and Engineering
University of Dundee

Student Assignment Brief

This document is for University of Dundee students for their use in completing assessed work. It should not be passed to third parties or posted on any other website.

Assignment Information	
Module Name:	Data Visualization Analytics
Module Code:	CS31001 / CS51007
Assignment Title:	Final Project
Assignment Weighting:	70%
Assignment Due Date	23.59 (11.59pm) on Sunday 7th December 2025

If you have questions about this assignment, please contact Dr Craig Ramsay (cdramsay@dundee.ac.uk)

Learning Outcomes

This assignment addresses the following learning outcomes of the module:

- Discuss data mining techniques
- Visualization of data
- Select data mining techniques
- Select appropriate data visualisations
- To develop and apply problem-solving, communication time-management, self-assessment and independent study skills

AI Usage Guidance

Given the broad and evolving nature of these tools, it is important to be clear about what is permitted in each assessment. Your module lecturers will indicate to what extent generative AI can be used. If you are unsure whether the use of a specific AI tool is permitted, please contact your module coordinator for guidance. Misuse or failure to disclose AI assistance may be treated as academic misconduct and subject to procedures outlined in the academic misconduct by students code of practice.

The levels of AI usage in assessment are:

RAG Rating	Permitted Level	Description
Red (Forbidden)	No	AI tools must not be used for this assessment. This includes generating content, conducting analysis, or providing answers to specific questions.
Amber (Restricted)	**Yes**	**AI tools may be used with caution, but their use must be explicitly acknowledged. For example, AI assistance in brainstorming, drafting, or structuring ideas is allowed, but the final submission must be your own work. Any AI-generated content must be fully reviewed, edited, and critically evaluated.**
Green (Permitted)	No	AI tools are permitted for this assessment without restriction, but their use must be explicitly acknowledged. You may use AI systems for research assistance, content generation, and other purposes, provided that the usage aligns with the learning outcomes and your

		understanding is demonstrated. You should engage critically with output and apply your own judgement.

Assignment Task

The objective of this final project is to demonstrate your comprehensive understanding of the data mining lifecycle by applying various techniques—clustering, classification, and regression—to a single dataset. You must open, prepare, analyse, model, and interpret the data, providing clear commentary and justifications at every stage of the process.

Your submission will be a single, well-structured Jupyter Notebook (.ipynb). The notebook must be fully runnable and include:

1. Code: Clean, well-commented code for every step.
2. Commentary & Observations: Detailed markdown cells with observations, interpretations, and justifications for the decisions you make (e.g., why you chose a particular k, why you scaled the data in a certain way, what the confusion matrix reveals).
3. Visualisations: Relevant plots for EDA, clustering evaluation, and model diagnostics.

The Dataset: Scottish Haggis

Imagine the following scenario. You have been commissioned by Lord Ramsay McCraig – a famous Scottish wildlife pioneer and a landowner of many vast estates throughout Scotland. Lord McCraig has been monitoring the population of one of Scotland's most famous wildlife creatures – the 'Haggis'. Rare sightings of the animal began a number of years ago in some of the Scottish Islands (the Islands of Iona, Skye, and Shetland). Since then, wildlife experts have managed to monitor the species. Three species of Haggis have been identified – the 'Wild Rambler' species, the 'Macduff' species, and the 'Bog Sniffler'. Wildlife volunteers have managed to record sightings of these animals, recording which Island they were sighted on, which year the sighting occurred, which species it was, and various measurements of the animal's features. The following measurements have been recorded for each haggis found: nose length, eye size, tail length, body mass, and gender.

You have been provided with a dataset that contains recordings of 344 haggis sightings. Your task is to explore this data and to apply a range of machine learning techniques.

Project Stages and Requirements

Stage 1: Exploratory Data Analysis (EDA)

This stage involves preparation tasks that will be necessary for subsequent modeling stages. Typical steps required here are as follows:

- **Data Loading & Initial Inspection:** Load the data and inspect the first few rows, data types, and summary statistics.
- **Exploratory Data Analysis (EDA):** Visualise the data (histograms, scatter plots, correlation matrices, etc.) to understand the distributions, relationships between features, and identify potential issues. **Provide commentary on what the data shows.**
- **Data Cleaning:** Handle missing values, incorrect data types, or identified outliers. Justify your cleaning decisions.
- **Feature Engineering:** [EXTRA / OPTIONAL]. Identify if there are any new features you can engineer from the data, if you think they will be useful for your analyses.
- **Scaling/Normalisation and/or Encoding:** Consider and discuss the need for scaling/normalisation and encoding for certain features in the data. Remember that some machine learning algorithms may or may not need scaling. Also, the stage at which you scale or encode data may depend on the algorithm, e.g., whether it should be done before or after a train-test split. Therefore, any scaling or encoding you apply to the data for earlier analysis may need to be reconsidered again for later analyses.

Stage 2: Unsupervised Learning (Clustering)
Use the prepared data for clustering. Typical steps required here will be as follows:

- Determine a value for 'k'. Show how you determine this and explain why you select your chosen value.
- Apply the k-means algorithm to the data, considering which features in the data you will be used, and considering any scaling or encoding required where necessary.
- Evaluate and interpret the clusters:
    - Perform an analysis of the clusters found. This involves looking at summary statistics and visualising the clusters across the original features.
    - Commentary: Describe the characteristics of each cluster and provide an interpretation of what each cluster represents in the context of the dataset.
- **Extra Credit:** Consider using a density-based clustering algorithm such as **DBScan** and comment on the differences in the clusters found, if any.

Stage 3: Supervised Learning (Classification)
Starting with Decision Trees, you will employ classification techniques to the data. Typical steps required are as follows:

- Doing a train-test split on the data, considering which features in the data you will be used, and considering any scaling or encoding required where necessary.
- Create a Decision Tree and fit it to the data.
- Visualise the tree to see what it shows, and comment on anything obvious.
- Evaluate the model, e.g., using metrics such as the following: Accuracy Score, Confusion Matrix, Classification Report (Precision, Recall, F1-Score). Provide a thorough **commentary** on the results, noting where the model performs well and where it struggles.
- Examine and comment on the **Feature Importances** derived from the Decision Tree model.
- **Extra Credit:**
    - Experiment with **hyperparameter modification** (e.g., max_depth, min_samples_split) to pre-prune the tree and comment on how it affects the accuracy/overfitting.
    - Consider post-pruning methods such as cost-complexity pruning.
    - Implement an **Ensemble Method** such as **Random Forests** and/or **XGBoost** to compare its performance against the single Decision Tree.

Stage 4: Supervised Learning (Classification Continued – Comparative Analysis)
Apply two additional classification methods to the *same* train/test split data from Stage 3 to compare and contrast them.

- **K-Nearest Neighbors (KNN):** Implement the KNN classifier. Show and explain how you determined a value for 'k'.
- **Logistic Regression:** Implement the Logistic Regression classifier.
- **Evaluation and Comparison:** Evaluate both models using the same metrics as Stage 3 (Accuracy, Confusion Matrix, Classification Report).
- **Logistic Regression Interpretation:** For Logistic Regression, examine the **coefficients** to determine which features have the strongest positive and negative influence on the target variable. Comment on these findings.
- **Conclusion:** Compare the performance of Decision Tree, KNN, and Logistic Regression and discuss which method is most effective for this dataset and why.

Stage 5: Supervised Learning (Regression)
Select feature(s) in your dataset to use for linear regression, including the 'target' you are trying to predict. Typical steps may be as follows:

- Create a new train-test split based on the features you are using. Perform necessary scaling/encoding (if required).
- Implement the Linear Regression model.

- Evaluate the model performance using standard regression metrics, e.g., R^2 score, mean absolute error, mean squared error or root mean squared error.
- Interpret the results and discuss the fit of the model to the data.

Other / extra considerations

The number of features in this dataset is relatively small. However, if there are occasions where the number of features / dimensions becomes larger after stages of encoding, you could consider the use of **Principal Component Analysis (PCA)** to reduce dimensionality, and comment on the impact it has to the outcomes of the machine learning methods. PCA can also be useful if trying to visualise high-dimensional data (e.g., in the form of a scatter plot, or similar).

All of the above should be encapsulated as steps and stages as within a single Jupyter Notebook. You can create the notebook using whatever tool you prefer, e.g., Jupyter, Colab, VS Code. You should ensure that your notebook is well structured and utilises 'markdown' / sections of text to create headings and sub-headings for each stage in the process. You should ensure that you provide comments on what you are doing and any key assumptions you are making as you work through the stages. You should ensure that you provide analysis of the outcomes that you present as you go along.

Deliverables

Please submit a single Jupyter Notebook file (a .ipynb file) that contains your data mining investigation.

How to Submit

Assignments should be submitted to My Dundee in the 'Assessment and Feedback' area.

Assignment Feedback

Feedback on your assessment will be provided via **My Dundee** and will be available within **three weeks** of the submission deadline. Any alteration to this will be communicated with you in advance through My Dundee.

Feedback will be provided in the form written or oral comments, and a grade.

Academic Integrity

By submitting this assignment, you are bound by the University of Dundee policies in relation to Academic Standards and Integrity. Please refer to Policy Statements, Regulations and Guidance for more information

Late Penalties

One grade point per day late (meaning if a submission is one day late and marked as a C2 it will receive a C3 grade). A day is defined as each 24-hour period following the submission deadline including weekends and holidays. Assessments submitted more than 5 days after the agreed deadline will receive a zero mark (AB).

Marking Criteria

See over page for landscape view of criteria

School of Science and Engineering
University of Dundee

Grade Band	1. Data Preparation & EDA (15%)	2. Clustering (K-Means) (10%)	3. Classification (Decision Trees) (15%)	4. Comparative Classification (KNN & LR) (15%)	5. Regression (Linear Regression) (15%)	6. Analysis, Commentary & Justification (20%)	7. Code Quality & Notebook Structure (10%)
A	Exemplary Data Prep. Cleaning, encoding, and scaling are optimal and critically justified. EDA visualisations are insightful and clearly structured to guide later modelling decisions.	All B criteria met, PLUS the notebook demonstrates enhanced aspects, e.g., implementing PCA and/or DBScan to provide comparative analysis and deeper insight into the data's structure.	All B criteria met, PLUS the student attempts advanced techniques like Random Forests (Ensemble Method) or demonstrates Hyperparameter Tuning/Post-Pruning to optimise performance.	Deep Comparative Analysis. The comparison is critical and well-supported by evidence. The interpretation of Logistic Regression coefficients is insightful and used to draw high-level conclusions about the data's inherent predictability.	All B criteria met, PLUS the student provides a critical discussion of the model's assumptions based on features selected, EDA, and/or further diagnostics.	Commentary is insightful, critical, and original. The student links findings across different stages to build a compelling narrative about the data.	Professional Standard. The notebook serves as a publishable, stand-alone report, easy for any professional to follow and reproduce.
B	Data preparation is robust and effective. Scaling/encoding choices are appropriate for the methods used. EDA is comprehensive, revealing key patterns and relationships.	K-Means is implemented with both Elbow and Silhouette used to justify k. Cluster analysis is detailed, focusing on key differentiating features and interpreting the clusters in the data's context.	Model evaluation is thorough and the commentary on the results is analytical. Feature Importances are correctly identified and interpreted, linking them back to the original data.	The comparison across Decision Tree, KNN, and Logistic Regression is detailed and comparative. Logistic Regression coefficients are correctly extracted and interpreted to understand feature influence on the target.	Regression metrics are interpreted correctly, and the commentary discusses the quality of the model fit and potential limitations.	Commentary and observations are consistently clear, analytical, and insightful at every major step. Decisions are well-justified based on evidence from the data or literature.	Code is efficient, elegant, and thoroughly commented. The notebook is professionally structured, using markdown cells effectively to deliver a seamless, narrative-driven report.
C	Data is correctly loaded, cleaned, and features are appropriately encoded and scaled. EDA	K-Means is implemented correctly. Either the Elbow Method or Silhouette Score is	Decision Tree is implemented with correct train-test split and data transformation. Model	Both KNN and Logistic Regression are implemented and evaluated using the standard set of	Linear Regression is implemented correctly with a suitable numerical target. Evaluated	Commentary is present at each stage, detailing what the data shows and the	Code is generally clean and well-commented. The notebook is logically structured

Grade Band	1. Data Preparation & EDA (15%)	2. Clustering (K-Means) (10%)	3. Classification (Decision Trees) (15%)	4. Comparative Classification (KNN & LR) (15%)	5. Regression (Linear Regression) (15%)	6. Analysis, Commentary & Justification (20%)	7. Code Quality & Notebook Structure (10%)
	includes basic visualisations (histograms, box plots).	used to determine k. Summary stats or basic plots are used to characterise the clusters.	is evaluated using Accuracy, Confusion Matrix, and Classification Report. Basic commentary provided.	metrics. A basic comparison of the three models is presented.	using R2, MAE, and MSE/RMSE.	decisions made. Interpretation of basic results (e.g., accuracy) is mostly correct.	with markdown headers separating the stages.
D	Basic loading and inspection is done. Initial steps to address missing values and categorical data attempted but may contain errors. Minimal EDA.	K-Means is attempted, but optimal k is arbitrarily chosen. Clusters are labelled but no meaningful analysis is performed.	Decision Tree model is implemented but model evaluation is limited (e.g., only accuracy score). Scaling/encoding is incorrectly applied (e.g., pre-split).	Both KNN and Logistic Regression are implemented, but evaluation is minimal, or the comparison is absent.	Linear Regression is attempted, but one or more required evaluation metrics are missing.	Commentary is descriptive but not analytical. Decisions (e.g., choice of k or split ratio) are stated but not justified.	Code is functional but minimally commented. The notebook is linear but lacks clear section headings.
Fail	Data is not loaded, or basic data types/missing values are not addressed. No EDA performed.	K-Means is not attempted, or the code fails to run. No attempt to find optimal K.	Train-test split is incorrect or missing. Decision Tree model not implemented.	Only one of the additional classifiers is attempted, or results are not evaluated.	Not attempted or code is fundamentally flawed.	Missing or only one-sentence descriptions of the code.	Code is messy, poorly commented, or notebook structure is confusing and hard to follow.