



SCHOOL OF SCIENCE AND TECHNOLOGY
Semester 1, 2026

FINAL EXAMINATION PAPER

DSC401: DATA SCIENCE IN PRACTICE

Important Instructions to Students:

1. Read the questions, instructions carefully and answer the questions.
2. You are allowed an **extra 10 minutes reading time** (during this time, you are not allowed to write with pen) and **180 minutes (3 Hours) to complete this paper**.
3. Silent, non-programmable calculators are allowed, but not supplied.
4. Write all your answers in the answer booklet provided. If you need extra sheets, please ask from the supervisor. Write your name, course code and student identification number in the answer booklet and in the attendance sheet.
5. This paper contains three (3) sections. All questions are compulsory to answer.
6. Write all your answers legibly. The marker has the right to award zero for answers that are not legible.
7. This is a **Closed Book Examination**. The University views plagiarism, cheating and dishonest practice as serious offences and if found, offenders are penalized.

Examination Paper Content (Format):

<u>Section</u>	<u>Question</u>	<u>Content</u>	<u>Marks</u>	<u>Suggested Time</u>
A	1 -5	True/False	5	9 minutes
B	6- 10	Multiple Choice	5	9 minutes
C	11 -16	Case Studies	90	162 minutes
		TOTAL	100	180 Minutes

Section A: True/False (5 Marks)

1. Data wrangling involves cleaning, transforming, and preparing data for analysis. (True / False)
2. In R, the filter() function is commonly used to remove columns from a dataset. (True / False)
3. A positive correlation means that as one variable increases, the other variable tends to increase as well. (True / False)
4. Clustering is a supervised machine learning technique. (True / False)
5. Sampling variability refers to differences that naturally occur between different samples taken from the same population. (True / False)

Section B: Multiple Choice (5 Marks)

6. Which of the following is primarily used for combining datasets based on a common field?

 - A. mutate()
 - B. summarise()
 - C. join()
 - D. arrange()
7. Which measure is most affected by extreme outliers?

 - A. Median
 - B. Mean
 - C. Mode
 - D. Interquartile Range
8. Which visualization is most suitable for showing the relationship between two quantitative variables?

 - A. Histogram
 - B. Scatterplot
 - C. Bar chart
 - D. Pie chart
9. In clustering analysis, what is the purpose of the “elbow method”?

 - A. To remove outliers
 - B. To normalize data
 - C. To determine the optimal number of clusters
 - D. To calculate prediction accuracy
10. Which of the following best describes overfitting in predictive modelling?

 - A. The model performs well on both training and test data
 - B. The model is too simple to capture patterns
 - C. The model memorizes training data and performs poorly on new data
 - D. The model has missing values in the dataset

Section C: Case Studies (90 Marks)

Question 11: Case Study: Data Wrangling

(15 marks)

The Fiji Sugar Corporation (FSC) collects data from multiple systems and identifies issues:

- Duplicate farmer records
- Missing values in `sugar_content`
- Region recorded inconsistently (e.g., “Western”, “west”)
- Some records show **negative tonnes delivered**
- Data stored across multiple tables

FSC must clean and prepare data before analysis.

Tasks (3 marks each):

- Identify and explain **two key data wrangling steps** FSC must perform before analysis.
- Explain two methods FSC can use to handle **missing `sugar_content` values**, and when each is appropriate.
- Some records show **negative tonnes delivered**. Explain why this is a problem and how FSC should handle it.
- Explain why FSC must standardise the **region variable**, and what problems occur if this is not done.
- FSC needs to **combine** farmers, deliveries, and payments datasets. Explain why combining datasets is necessary and what risks exist if done incorrectly. (*Hint: Full picture vs errors.*)

Question 12. Case Study: Exploratory Data Analysis

(15 marks)

The Fiji Sugar Corporation (FSC) is analysing sugarcane delivery data from **3,200 farmers across Western, Northern, and Central regions**. Dataset variables include:

- *farmer_id*, *region*, *farm_size_ha*, *tonnes_delivered*, *sugar_content (%)*, and *rainfall_mm*

EDA results:

- Mean tonnes delivered = **72 tonnes**
- Median tonnes delivered = **60 tonnes**
- Standard deviation = **35 tonnes**
- Distribution is **right-skewed (long right tail)**
- ~5% of farmers produce **>180 tonnes**
- Boxplots show **extreme high-value outliers**
- Scatterplot shows **positive relationship (farm size vs production)**
- Correlation (rainfall vs production) = **0.45**
- Regional patterns:
 - Western → higher mean, higher variability
 - Northern → lower mean, more consistent
- When grouped by **burnt vs non-burnt cane**, relationships change

Tasks (3 marks each):

a) The mean tonnes delivered per farmer is **72 tonnes**, while the median is **60 tonnes**.

Explain what this difference between the **mean and median tonnes delivered** suggests about the **shape of the distribution (distribution type)** of sugarcane production across FSC farmers.

b) The standard deviation of **tonnes delivered per farmer** is **35 tonnes**.

What does this tell us about the **spread and consistency of sugarcane production across farmers**.

c) Approximately **5% of farmers deliver more than 180 tonnes of cane**, and boxplots confirm extreme high values.

Explain what these **high-production farmers (tonnes delivered >180)** likely represent in the FSC context and evaluate whether FSC should remove these high-production values from the dataset, considering both **statistical validity and real-world business meaning**.

d) Exploratory Data Analysis shows a moderate positive correlation ($r = 0.45$) between *rainfall* (*rainfall_mm*) and *tonnes delivered*.

Interpret and explain what FSC can reasonably conclude from this relationships, and what conclusions **cannot be made**, particularly in terms of causation.

e) When the data is split into **burnt cane vs non-burnt cane**, the relationship between **farm size (farm_size_ha)** and **tonnes delivered** changes significantly.

State the statistical concept (paradox) illustrated by the change in the relationship between **farm size and tonnes delivered** after splitting the data into burnt and non-burnt cane. Explain why this is important for FSC to avoid misleading conclusions when analysing **aggregated data versus subgroup data**. (*Hint: Overall trends may differ from subgroup trends.*)

Question 13. Case Study: Data Management

(15 marks)

The Fiji Sugar Corporation (FSC) stores operational data across multiple datasets, as follows:

- farmers.csv - (*farmer_id, region, farm_size_ha*)
- cane_deliveries.csv - (*delivery_id, farmer_id, tonnes_delivered, sugar_content*)
- payments.csv - (*farmer_id, total_proceeds, amount_paid, balance*)

During analysis, FSC identifies the following issues:

- Some farmers have **no delivery records**
- Some delivery records have **no matching farmer_id**
- Some payment records are **missing or incomplete**
- Analysts need to combine datasets to evaluate:
 - farmer productivity
 - fairness of payments
 - regional performance

FSC must prepare a **clean, integrated dataset** to support decision-making.

Tasks (3 marks each)

a) Explain why FSC cannot directly analyse raw datasets (farmers, deliveries, payments) without proper data management.

b) Which join should be used for the following:

- FSC wants to analyse **only farmers who have valid delivery records**.
- FSC wants to identify farmers who are registered but have **no cane deliveries**
- FSC finds some delivery records that **do not match** any farmer.
- Which join should be used if FSC wants to perform a **full** audit including:
 - all farmers
 - all deliveries
 - all mismatches

c) FSC groups data by **region** and calculates: *total tonnes delivered* and *average payment per tonne*. How would grouping data by region support decision-making at FSC? You can focus on the above two variables (total tonnes and average payment).

d) Explain how using incorrect joins or poorly managed data could lead to **wrong business decisions** at FSC.

Question 14: Machine Learning and Prediction (15 marks)

FSC wants to predict whether harvested sugarcane batches will be classified as:

- “*High Quality*” or
- “*Low Quality*”.

The dataset includes: *tonnes_delivered*, *sugar_content*, *burnt_flag*, *rainfall_mm*, *soil_quality*, *transport_delay_hours*

The data science team develops several models: *Logistic regression*, *Decision tree*, *Random forest*, and *Neural network*. The team observes:

- One decision tree performs extremely well on training data but poorly on testing data
- The random forest performs well on both training and testing data
- Missing diseased crops would create major economic losses

Tasks (3 marks each)

- a) Explain why this is a supervised learning problem.
- b) Using an FSC example differentiate between *regression* and *classification*.
- c) One model performs extremely well on training data but poorly on test data. Identify and explain this problem.
- d) Explain why random forests often perform better than a single decision tree. (*Hint: Combining multiple trees.*)
- e) FSC wants to detect diseased crops where missing diseased cases is dangerous. Which evaluation metric is most important and why? (*Hint: Think about false negatives.*)

Question 15: Supervised Learning and Modelling (15 marks)

FSC develops predictive models to estimate *tonnes_delivered* using:

- farm_size_ha
- rainfall_mm
- sugar_content
- burnt_flag

The analytics team develops:

- simple linear regression
- multiple regression
- decision tree model

The following observations are made:

- Simple regression has lower predictive accuracy
- Multiple regression performs better
- One model appears overfitted
- A statistical test comparing burnt vs non-burnt cane gives $p = 0.04$

Tasks (3 marks each)

a) Using the FSC scenario, explain the difference between the following.

- explanatory variables
- response variable

b) Explain why multiple regression is likely better than simple regression in this scenario. (*Hint: Multiple factors influence production.*)

c) FSC obtains $p = 0.04$ when comparing burnt vs non-burnt cane. Interpret this result using $\alpha = 0.05$. (*Hint: Statistical significance.*)

d) Explain why overfitting is dangerous for FSC forecasting systems.

e) Explain why test datasets are essential when evaluating predictive models.

Question 16: Unsupervised Learning and Clustering (15 marks)

FSC wants to better understand different types of sugarcane farmers across Fiji.

The dataset includes:

- farm_size_ha
- tonnes_delivered
- sugar_content
- rainfall_mm
- distance_to_mill
- fertilizer_usage

FSC applies:

- k-means clustering
- hierarchical clustering
- anomaly detection

The analytics team observes:

- Several dense farmer groups
- Some isolated farmers far from all clusters
- Different clustering results when variables are not scaled
- An elbow graph suggesting 3–4 clusters

Tasks (3 marks each)

- Explain why this is an **unsupervised** learning problem.
- Explain what dense farmer groups likely represent.
- Some farmers appear far away from all clusters. Explain what these farmers may represent and why FSC should investigate them.
- Explain why scaling variables is important before **k-means clustering**. (*Hint: Distance calculations.*)
- Explain how clustering results could help FSC operationally.

THE END