Subject Outline

Subject Name	Advanced Statistical Methods for Data Scientists
Subject Code	MA5821
Study Period	SP5 2019
Study Mode	External
Campus	JCU Online
Subject Coordinator	Neil Fraser

We acknowledge the Traditional Owners of the lands and waters where our University is located and actively seek to contribute and support the JCU Reconciliation Statement, which exemplifies respect for Australian Aboriginal and Torres Strait cultures, heritage, knowledge and the valuing of justice and equity for all Australians.

Cairns
Singapore
Townsville

Pre-requisites

Before attending this course, you should have taken a college-level course in statistics (e.g MA5820 Statistical Methods) that taught explanatory modelling, t-tests, ANOVA, distribution analysis, hypothesis testing, and regression techniques or have equivalent knowledge. Students should also have an understanding of the basics of computing, data management, different data types and some programming.

This subject outline has been prepared by Neil Fraser for the College of Science and Engineering, Division of DTES, James Cook University. Updated 28.02.2018.

Q1. This subject is offered across more than one campus and/or mode and/or teaching period within the one calendar year.	Yes ☒	No ☐
Q2. If Yes (Q1), the design of all offerings of this subject ensure the same learning outcomes and assessment types and weightings.	Yes ☒	No ☐
Q3. If no (Q2), **[Type here]** has authorised any variations, in terms of equivalence.		

Subject Outline Peer Reviewer

Name	Ron White
Position	Professor
Date Reviewed	

Staff Contact Details

Teaching Team	Staff Member	Room	Phone	Email	Consultation times*
Subject Coordinator	Neil Fraser	N/A			
Tutor 1	Banmali Pradhan			banmali.pradhan@jcu.edu.au	
Tutor 2	Mohammed Zilani			mohammed.zilani@jcu.edu.au	
Tutor 3					

*Other consultation times by appointment only.

Contents

1 Subject at a glance

1.1 Student participation requirements

The JCU Learning, Teaching and Assessment Policy (4.3) indicates that, "a **3 credit point subject** will require a **130 hour work load** of study-related participation including class attendance over the duration of the study period, **irrespective of mode of delivery"**. This work load comprises **timetabled hours** and **other attendance requirements**, as well as **personal study hours,** including completion of online learning activities and assessment requirements. Note that "attendance at specified classes will be a mandatory requirement for satisfactory completion of some subjects" (Learning, Teaching and Assessment Policy, 5.10); and that additional hours <u>may</u> be required per week for those students in need of **English language, numeracy** or **other learning support.**

Key subject activities	Time	Day & Date	Room/Location
e.g. Lecture (1 hour per week)	Refer to JCU Timetable 2019		
e.g. Tutorial (2 hours per week)	Refer to JCU Timetable 2019		
Other mandatory attendance requirements:	Online Tutorials	Wednesday Evening Friday Evening Sunday Morning	Online

For information regarding class registration, visit the Class Registration Schedule.

1.2 Key dates

Key dates	Date
Census date	12 September, 2019
Last date to withdraw without academic penalty	19 September,2019
Assessment 1 – Weekly quizzes. Total: 30% • Week 1 quiz: 10% • Week 3 quiz: 5% • Week 4 quiz: 5% • Week 5 quiz: 5% • Week 6 quiz: 5%	Week 1 quiz : A1A Due Sunday Week 1 Week 3 quiz : A1B Due Sunday Week 3 Week 4 quiz : A1C Due Sunday Week 4 Week 5 quiz : A1D Due Sunday Week 5 Week 6 quiz : A1E Due Sunday Week 6 See LearnJCU for details on date and time.
Assessment 2 – Weekly workbook exercise submissions (including short answers). Total: 30% • Week 2 submission (Regression) : 7.5% • Week 3 submission (General Linear Modelling) : 7.5% • Week 4 submission (Logistic regression) : 7.5%	• Week 2 submission: A2A Due Sunday of Week 2. • Week 3 submission: A2B Due Sunday of Week 3. • Week 4 submission: A2C Due Sunday of Week 4. • Week 5 submission: A2D Due Sunday of Week 5. See LearnJCU for details on date and time

Key dates	Date
• Week 5 submission (Decision Trees and Cluster Analysis) : 7.5%	
Assessment 3 – Capstone project. Total :40%	Released in Week 4. Due Wednesday of Week 7. See LearnJCU for details on date and time

2 Subject details

2.1 Subject description

To be effective an organisation needs to be able to build its data assets by translating raw operational data into strategic information for decisions, data collaborations, personalising services to customers, for making predictions and for forecasting. A data scientist is one of the key contributors to this type of activity and should be able to identify and work with a range of tools and data structures to ensure organisations stay effective and efficient in a competitive world. In this unit, you will be encouraged to learn skills required to succeed in today's highly analytical and data-driven economy using robust industry tested software for data science. This subject will introduce students to practical applications and concepts involved in advanced statistical modelling in SAS. Topics include

- linear modelling with multiple predictor variables that maybe continuous or categorical in nature;
- Conditional Probability and the odds ratio
- drawing inferences
- checking model diagnostics and model selection
- techniques for coping with data that are temporally or spatially correlated.

2.2 Subject learning outcomes

Students who successfully complete this subject will be able to:

1. demonstrate sound knowledge of the basic principles and theories that underpin advanced statistical modelling methods.
2. effectively integrate and execute advanced statistical modelling theories and processes in SAS software to solve authentic problems
3. retrieve, analyse, synthesise and evaluate outputs produced using advanced statistical modelling methods in SAS software.
4. Critically examine different approaches to advanced statistical problems

These outcomes will contribute to your overall achievement of **course learning outcomes.** Your course learning outcomes can be located in the entry for your course in the electronic JCU Course Handbook 2019 (see *Academic Requirements for Course Completion*).

2.3 Learning and teaching in this subject

Week 1 of this subject introduces students to fundamental concepts used to do predictive modelling and the environment used for SAS virtual laboratories. The entire week 2 is a recap on regression and an introduction to multiple linear regression with a tutorial in SAS Visual Analytics. Week 3 introduces the practice and use of general linear modelling. Week 4 examines and explains logistic regression with a tutorial in SAS Visual Analytics. Week 5 cover Decision trees and cluster analysis using SAS Visual Analytics. Week 6 introduces forecasting and time series analysis with a final discussion model selection and diagnostics. .

2.4   Student feedback on subject

As part of our commitment at JCU to improving the quality of our courses and teaching, we regularly seek feedback on your learning experiences. Student feedback informs evaluation of subject and teaching strengths and areas that may need refinement or change. ***YourJCU Subject and Teaching Surveys*** provide a formal and confidential method for you to provide feedback about your subjects and the staff members teaching within them. You will receive an email invitation when the survey opens. We value your feedback and ask that you provide constructive feedback about your learning experiences for each of your subjects, in accordance with responsibilities outlined in the Student Charter. Refrain from providing personal feedback on topics that do not affect your learning experiences. Malicious comments about staff are deemed unacceptable by the University.

2.5   Subject resources and special requirements

1) VisualAnalytics / Statistics getting started guide on TUN
Chapter 1 About Exploration, Chapter 2 Explore a Data Source , Chapter 3 Cheat Sheet for Explorations
2) ExploringDatawithSASVisualAnalytics.pdf   Chapter 1
3) Statistics1-IntroductiontoANOVARegressionandLogisticRegression.pdf Chapter 1 Quick Review of Statistical Concepts (pg33-46, pg53-56) ; Chapter 2 ANOVA -Regression (pg 1-8, 53-56)

SAS Data and Statistics References
SAS Data Dictionaries on Teradata University Network
SAS Glossary of Definitions

3 Assessment details

3.1 Requirements for successful completion of subject

In order to pass this subject, you must:
• Achieve an overall percentage of 50% or more;

Assessment items and final grades will be reviewed through moderation processes (Learning, Teaching and Assessment Policy, 5.13-5.18). It is important to be aware that assessment "is always subject to final ratification following the examination period and that no single result represents a final grade in a subject" (Learning, Teaching and Assessment Policy, 5.22.).

3.1.1 Inherent requirements

Inherent requirements are the fundamental abilities, attributes, skills and behaviours needed to achieve the learning outcomes of a course while preserving the academic integrity of the university's learning, assessment and accreditation processes. Students and prospective students must be able to demonstrate that they have acquired or have the ability to acquire the inherent requirements for their degree.

Reasonable adjustments may be made to assist students manage additional circumstances impacting on their studies provided these do not change the academic integrity of a degree. Reasonable adjustments do not alter the need to be able to demonstrate the inherent requirements of the course. Students who believe they will experience challenges completing their degree or course because of their disability, health condition or other reason should discuss their concerns with an accessibility Services team member or a member of College staff, such as the Course Coordinator. In the case where it is determined that inherent requirements cannot be met with reasonable adjustments, the University staff can provide guidance regarding other study options.

3.2 Feedback on student learning

Feedback for students will be provided on some assessment items.

3.3 Assessment tasks

ASSESSMENT 1: FIVE QUIZZES

Aligned subject learning outcomes	1,2,3
Aligned professional standards/ competencies	JCU-SAS Joint Certificate in Advanced Data Science SAS Certified Statistical Business Analyst Using SAS 9: Regression and Modelling
Group or individual	Individual
Weighting	30%
Due date	Sunday Weeks 1,3,4,5 and 6

ASSESSMENT TASK 1: DESCRIPTION

- Short Quiz : Correct Answer

ASSESSMENT TASK 1: CRITERIA SHEET

All assessment questions will be marked by comparing student answers to a model set of solutions and marking scheme prepared by the lecturer.

ASSESSMENT 2: SHORT ANSWER AND WORKBOOK

Aligned subject learning outcomes	☐ 1, 2, 3, 4
Aligned professional standards/ competencies	JCU-SAS Joint Certificate in Advanced Data Science SAS Certified Statistical Business Analyst Using SAS 9: Regression and Modeling
Group or individual	Individual
Weighting	30%
Due date	Sunday Weeks 2 to 5

ASSESSMENT TASK 2: DESCRIPTION

- SAS Visual Analytics Workbook Submission with answers

ASSESSMENT TASK 2: CRITERIA SHEET

All assessment questions will be marked by comparing student answers to a model set of solutions and marking scheme prepared by the lecturer.

ASSESSMENT 3: CAPSTONE PROJECT

Aligned subject learning outcomes	☐ 1,2,3
Aligned professional standards/ competencies	JCU-SAS Joint Certificate in Advanced Data Science SAS Certified Statistical Business Analyst Using SAS 9: Regression and Modeling
Group or individual	Individual
Weighting	40%
Due date	Wednesday of Week 7

ASSESSMENT TASK 3: DESCRIPTION

- Written submission (10 pages) and workbook submission

ASSESSMENT TASK 3: CRITERIA SHEET

This capstone project assessment involves writing a report that summarises a data science related investigation that you have selected from one of the datasets listed in this document and available in the Teradata University Network SAS Visual Analytics virtual computing environment.

The investigation must involve:

1. An initial data exploration and explanation of the SAS Data sets listed in the assessment capstone criteria document. These are available online in the Teradata University Network and have available metadata (see link). Please note that any data set used in the previous weeks for tutorials or assignments are not to be used in this assignment (e.g. BIGPVA, PVA_DATA, insights toys, CARS, bio-organics).

2. A hypothesis that you wish to test based on your initial data selection from point 1 above and the modelling techniques to be deployed.

3. Modelling, statistical output and any model comparison for optimisation with SAS Visual Analytics completed.

4. A final written submission in word and evidence of the steps taken using SAS Visual Analytics to reach your conclusions (e.g model visualisations, summary stats tables etc..) and insights generated.

The written summary should cover the pre-processing/exploratory steps carried out, the statistical models evaluated and finally the comparison you have made. These are independent choices and decisions that require full explanation of your analysis process and the hypothesis that you have tested. Please include relevant screenshots of all the charts or model diagnostics that you created and any calculated fields, parameters used during model selection and evaluation.

The report should not exceed 10 pages and be readable with the following:

- Effective use of hypothesis, analytical approach and a conclusion.

- Contains relevant screen shot(s) showing the visualisation used from SAS Visual Analytic

- Graduate Standard of academic writing (structure, reference, grammar etc.)

The capstone assessment is to show you have grasped the important concepts and techniques associated with statistical modelling, model selection and evaluation using a coherent written analysis of a clear hypothesis. While marks are not awarded for neatness, students may be penalised for poorly written or extremely untidy work.

4 Submission and return of assessment

4.1 Submission of assessment

- All assessments are submitted through Learn Ultra.

- A cover sheet with student name and id should be attached to the assignment

- In the case where a disruption to studies application is approved, the student may be offered an alternative assessment or may receive a mark based on the percentage mark achieved by the student in one or more other assessment tasks, at the unit convenor's discretion.

Note that the Learning, Teaching and Assessment Policy (5.22.3) outlines a uniform formula of penalties that will be imposed for submission of an assessment task after the due date. **This formula is 5% of the total possible marks for the assessment item per day including part-days, weekends and public holidays**. After 20 days, the assessment item thus would be awarded 0 marks (i.e. 5% x 20 = 100% of total possible marks in penalties).

4.2 Return of assessment

Feedback on marked assessments will be available in the Gradebook in Learn Ultra.

It is the responsibility of students to view their marks for each within-session assessment on Learn JCU within 20 working days of posting. If there are any discrepancies, students must contact the unit convenor immediately. Failure to do so will mean that queries received after the release of final results regarding assessment marks (not including the final exam mark) will not be addressed.

Please see The Learning Centre website for other important student information pertaining to plagiarism and referencing, examinations advice and student support services.

5 Subject calendar MA5821 Advanced Statistical Methods for Data Scientists

Please note, the sequence of some topics may change due to staff availability, resourcing, or due to unforeseen circumstances.

Week/Date/Module		Topics Covered	e.g. Tutorial	Readings/Preparation	Relationship to Assessment
O	Orientation	Introduction to SAS® software and training profiles	Register for Teradata University Network	Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression Free Elearning	VLE environment setup
1	Introduction to SAS VA	Topic 1: Why are analytics and statistics important? Topic 2: Introduction to core Concepts in Advanced Statistical modelling Topic 3 Advanced Probability Topic 4: Frequentist vs Bayesian schools	Exploring SAS Visual Analytics user interface – An example	- SAS(R) Visual Analytics/ Visual Statistics 7.2: User's Guide - "The Four Types of Estimable Functions," in the SAS/STAT® 13.2 User's Guide.	Introduction to Assignment Submission format Quiz A1A
2	SAS Regression	Topic 1: Regression Recap Topic 2: Mutliple Linear Regression Topic 3: Running Regression Models in SAS Visual Analytics /Statistics	Self-Learning practical: Building a Linear Regression (PVA_DATA_Demo)	Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression Statistics 2: ANOVA and Regression	Workbook A2A
3	General Linear Modelling	Topic 1 : Regularization	Tutorial 1: General linear model properties		Quiz A1B Workbook A2B

Week/Date/Module	Topics Covered	e.g. Tutorial	Readings/Preparation	Relationship to Assessment
	Topic 2 : missing Values in Regression Modelling Topic 3: Generalized Linear Model (GLM)	Tutorial 2: Analysing general linear model results Self-Learning Practical: Building a GLM Model		
4	Predictive Modelling 2 Classification Models	Topic 1: Introduction to Logistic Regression Topic 2: The Logistic Function Topic 3: Predictive Modelling using Logistic Regression in SAS Visual Analytics	Self-Learning Practical: Building a Logistic Regression Model in SAS Visual Analytics	Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression Statistics 2: ANOVA and Regression	Quiz A1C Workbook A2C
5	Predictive Modelling 3 Decision Trees and Clustering	Topic 1: Introduction to Decision Trees Topic 2: Decision Trees using SAS Visual Analytics / Statistics Topic 3: Cluster Analysis	Tutorial Walkthrough: Building a Decision Tree Self-learning practical: Building a Decision Trees with PVA data		Quiz A1D Workbook A2D

Week/Date/Module	Topics Covered	e.g. Tutorial	Readings/Preparation	Relationship to Assessment
		Self-learning practical: Building a Cluster Model		
6	Time Series and Model Comparison	Topic 1: Time Series analysis Topic 2: Model Selection & Assessment	Self-paced Practical exercise: Business forecasting in SAS Visual Analytics 1.Self-Learning Practical: Logistic Regression vs GLM Model 2.Self-Learning Practical: Logistic Regression vs Decision Tree	Quiz A1E
7				A3 capstone Project