Assessment Brief and Marking criteria 2021-22

Module title	**Principles of Data Science**
CRN	G340 M0001 / CRN41140
Level	7
Assessment title	**Statistical Analysis and Dashboard Creation**
Weighting within module	This assessment is worth 100% of the overall module mark.
Submission deadline date and time	**10/01/2022 at 4:00 pm.**

Module Leader/Assessment set by
Safwan Umer, Newton 225, ext: 55053; email: s.umer1@salford.ac.uk

How to submit
Your assignment should be submitted through blackboard and should be in the form of a single zipped file containing your report and any associated material (code, for example). Please check that the zip file:

1. Is valid and openable
2. Contains the full MS word document containing your work
3. Contains any additional material is clearly labelled sub-directories with a README.TXT document in the parent directory explaining the content and its relevance to the assignment. If this is an application/code, a fully working version should be included with a clearly written description of the application and its use.
4. Has a name that uniquely identifies this as your work.

Assessment task details and instructions

Your task is to demonstrate your newly developed knowledge and understanding of data handling, validation, anonymisation, ethics, statistical analysis, and visualisation by exploring and presenting data from an extensive, complex data set.

World Development Indicators (WDI) are the primary World Bank collection of development indicators, compiled from officially recognized international sources. WDIs present the most current and accurate global development data available, and includes national, regional and global estimates. The dataset for this assignment can be downloaded from https://databank.worldbank.org/source/world-development-indicators that encapsulates world indicators data for counties around the world, decomposed into categories of activity and metrics. You also have yearly snapshot data for these countries that could be construed as time series. Download the dataset by selecting number of countries and variables you want to work with.

This assignment proposes a considered and justifiable scheme for an information dashboard to present a country's economic health within an information-rich, intuitively comprehensive, single screen dashboard format. Your reasoned thinking, research, and critical evaluation of both the problem resolution and proposed solution form a substantive part of this work and will present the rationale for the proposal in the form of a report. A data set comprising wide-ranging economic data for many countries is provided for exploration, evaluation, and experimentation to justify the approaches and decisions made.

This assessment requires a complete data analysis, and a working dashboard prototype to be presented. It should make a feasible and justifiable, worked schema that demonstrates the critical evaluation of the processes of data preparation, validation, anonymization, ethics, algorithmic fairness, analysis and/or modelling and prediction and justifiable research into composition, layout, function, and form**.** Justification of the approaches taken for statistical analysis and visualisation is expected and worked examples should be provided.

The requirements for the proposed dashboard are:

1. A single-screen presentation of at least 10 countries' economic health and the profile of their economic activity profile.
2. Clear, effective representational presentation of all factors in a coherent, intuitively comprehensive form.
3. A schema applicable to the full range of countries presented in the dataset without modification to the dashboard form or structure. (i.e., the dashboard should support a side-by-side comparison of multiple countries and/or financial years).
4. Presentation of future prediction/trend for economic profile based on historical data.
5. Relational modelling showing relative performance against stochastically defined groups of countries within the data set.

The requirements for the proposed statistical analysis are:

1. Define a research objective based on the dataset. For instance, to compare the trade situation of the least developed countries with developed countries.
2. Based on the objective, select at least 10 suitable countries of your choice.
3. Choose a set of indicators according to the objective with at least 10 years of data.
4. Start to complete the following tasks. Also, present and interpret your findings and results in the report as much as you can and show the thorough SAS analytics steps.
    4.1. Do a comprehensive descriptive statistical analysis (e.g., Mean, Median, Mode, Standard deviation, Skewness and Kurtosis) on the data.

4.2. Do a correlation analysis for the indicators as much as you think is enough for the defined objective.
4.3. Do a regression analysis. Explain why the selected regression technique is better for the defined objective and show if you've found any similar research in the literature.
4.4. As a researcher, do a comparative analysis of the hypothesis testing approaches and explain when and why you use them? Then define two hypotheses related to the objective and test them.

5. In general, describe the steps that you've taken for data preparation, outlier detection, dealing with missing data, and data privacy protection.

You must use SAS programming for the statistical analysis part. You can use Tableau, Power BI or MS Excel technologies/applications to experiment with analysis and develop the dashboard (other tools could be discussed with the instructor). The submission of the assignment will be in the form of a report (guideline 40 pages) that presents the proposal, explains the rationale for the approaches used to analyse and display the data components, and critically evaluates subject domain research (data ethics and data visualisation) and the final implemented prototype.

Assessed intended learning outcomes

On successful completion of this assessment, you will be able to:

A- Knowledge and Understanding

1. Analyse a data science project to devise a structure for its implementation, analysis, and evaluation, justifying any decisions made.
2. Critically assess the relative strengths and uses of a range of statistical analysis techniques (including t-tests, ANOVA, linear regression, multiple regression models and categorical data analysis, test of hypothesis).
3. Present and visualise the statistical results, analysing key findings.
4. Evaluate the quality of graphs according to their expressiveness and effectiveness.

B- Practical, Professional or Subject Specific Skills

1. Understand the history and context of data science ethics, skills, challenges, and methodologies the term implies.
2. Will learn how to work with a dataset that possibly is not in your domain expertise, and you don't have prior knowledge and understanding of that field.
3. Develop skills in presenting quantitative data using appropriate displays, tabulations, and summaries.
4. Understand the nature of sampling variation and the role of statistical methods in developing and testing hypotheses.
5. Select and use appropriate statistical/ML methods in the analysis of complex datasets.
6. Present findings based on statistical analysis in a clear, concise, and understandable manner.
7. Select the proper visualization methods for a given data analysis and presentation problem.

C- Transferable Skills and other Attributes

1. Technical report writing.
2. Ability to use tools and techniques for statistical analysis.
3. Presenting data in a manner accessible to non-technical stakeholders.
4. Data Science Ethics, Information governance, information Literacy and Data Protection

Module Aims
The module is focused on the underpinning knowledge and practical skills needed for working within the data sciences industry.

Word count/ duration (if applicable)
Your assessment should be about 4000 words (20 pages).

Feedback arrangements

You can expect to receive individual feedback in the form of an annotated marking matrix with specific comments for each section, general comments for the work and up to 3 specific areas for improvement.

Support arrangements
You can obtain support for this assessment by contacting Safwan Umer for the technical aspects of the module. Further support can be obtained from the university as follows:

askUS
The University offers a range of support services for students through askUS.

Good Academic Conduct and Academic Misconduct
Students are expected to learn and demonstrate skills associated with good academic conduct (academic integrity). Good academic conduct includes the use of clear and correct referencing of source materials. Here is a link to where you can find out more about the skills which students require http://www.salford.ac.uk/skills-for-learning.
Academic Misconduct is an action which may give you an unfair advantage in your academic work. This includes plagiarism, asking someone else to write your assessment for you or taking notes into an exam. The University takes all forms of academic misconduct seriously. You can find out how to avoid academic misconduct here https://www.salford.ac.uk/skills-for-learning.

Assessment Information
If you have any questions about assessment rules, you can find out more here.

Personal Mitigating Circumstances
If personal mitigating circumstances may have affected your ability to complete this assessment, you can find more information about the personal mitigating circumstances procedure here.

Personal Tutor/Student Progression Administrator
If you have any concerns about your studies, contact your Personal Tutor or your Student Progression Administrator.

Assessment Criteria
It would be best to look at the assessment criteria to determine what we are explicitly looking at during the assessment.

In Year Retrieval Scheme
Your assessment is not eligible for in year retrieval. If you are eligible for this scheme, you will be contacted shortly after the feedback deadline.

Reassessment

If you fail your assessment and are eligible for reassessment, you will be allowed to re-do the assignment based on the feedback given. The submission for this will be at the 3rd-trimester re-submission date (usually the end of August).

Assessment criteria

	0-9%	10-19%	20-29%	30-39%	40-49%	50-59%	60-69%	70-79%	80-89%	90-100%
Level 7 Assessment Scale	Extremely Poor	Very poor	Poor	Inadequate	Unsatisfactory	Satisfactory	Good	Very Good	Excellent	Outstanding
Statistical Analysis	Some analyses based on an invalid data set has been done. None of the defined tasks has been done correctly, even according to the minimum expectations.		The objective of the research has not been defined. Less than minimum sample size has been used, and the approach for selecting this sample has not been justified. There is a poor consideration of data preparation. It means most of the data has been used in a raw format. Very little descriptive statistical analysis has been done. Poor correlation analysis has been done.		The objective of the research has been defined, but it is not explained well. A minimum sample size has been used, and the approach for selecting this sample has not been justified. A fair consideration of data preparation is included where there is a primary scheme shows how part of the general principles and approaches have been used. Little descriptive statistical analysis has been done. A satisfactory correlation analysis has been done.		The objective of the research has been defined and is adequately explained. A minimum sample size has been used, and the approach for selecting this sample has been justified. There is a reasonable consideration of data preparation in place. That is a thorough proposal that shows how general principles and approaches have been used. Somewhat descriptive statistical analysis has been done. Good correlation analysis has been done.		The objective of the research has been defined and is adequately explained. More than minimum sample size has been used, and the approach for selecting this sample has been justified. Advanced consideration of data preparation. Whereas besides general preparation, handling missing data and outlier data detection algorithms have been utilized. To a great extent descriptive statistical analysis has been done. Excellent correlation analysis has been done. A suitable regression analysis method has been chosen,	

	0-9%	10-19%	20-29%	30-39%	40-49%	50-59%	60-69%	70-79%	80-89%	90-100%
Level 7 Assessment Scale	Extremely Poor	Very poor	Poor	Inadequate	Unsatisfactory	Satisfactory	Good	Very Good	Excellent	Outstanding
			A poor regression analysis has been done. The method is not adequate, and a significant result has not been obtained. Preliminary comparative analysis of the hypothesis testing and part of the requested test of hypothesis has been included. The report consists of a minimum explanation of the results without demonstrating SAS analytics steps.		An acceptable regression analysis method has been chosen, but the results are not described. Fair comparative analysis of the hypothesis testing and part of the requested test of hypothesis has been included. The report consists of a reasonable explanation of the results and a partial demonstrating SAS analytics step.		A good regression analysis has been chosen, and the results are justified. Good comparative analysis of the hypothesis testing and requested test of hypothesis has been included. The report consists of an elaborate explanation of the results and demonstrating comprehensively SAS analytics steps.		results are justifiable, and some predictions have been made based on the model. Comparative analysis of the hypothesis testing and requested test of hypothesis has been included. The report consists of highly precise and well-explained results and demonstrating comprehensively SAS analytics steps.	
Data dashboard: *Visual composition,*		No use of data science ethics.	Functional data representations using basic forms without modification		A consistent but basic proposal that shows how general principles and approaches have been used to define a coherent presentation		Refined representation of data using tailored representational forms that extend and refine the basic offering of the packages considered.		Comprehensive representation of data using tailored representational forms that extend and refine the	

	0-9%	10-19%	20-29%	30-39%	40-49%	50-59%	60-69%	70-79%	80-89%	90-100%
Level 7 Assessment Scale	Extremely Poor	Very poor	Poor	Inadequate	Unsatisfactory	Satisfactory	Good	Very Good	Excellent	Outstanding
form and structure			Single country case considered and presented as a worked example		Functional representation of raw data based on standard representations with minimal modification of the attributes to reflect basic perceptual and cognitive features Indirect representation of data with minimal analysis or pre-preparation Some justification of approaches and principles applied		Functional representation of data based on standard representations with minimal modification of the attributes to reflect comprehensive perceptual and cognitive features Consideration of comparative analysis presented within the visual representation for a small selection of countries Consideration of a common perceptual model and justification of this in a cognitive context.		basic offering of the packages considered. Comprehensive practice in visualisation techniques, ethics and information assurance Consideration of comparative analysis presented within the visual representation all data ranges and scales Expanded definition of a common perceptual model and justification of this in a cognitive context. Detailed justification of approach taken, principles adopted, and assumptions made about the representational paradigm.	

	0-9%	10-19%	20-29%	30-39%	40-49%	50-59%	60-69%	70-79%	80-89%	90-100%
Level 7 Assessment Scale	Extremely Poor	Very poor	Poor	Inadequate	Unsatisfactory	Satisfactory	Good	Very Good	Excellent	Outstanding
									Detailed justification and application of data ethics principles base on UK data ethics framework.	
Data dashboard: *Proposal*	None		Inconstant/uncoherent proposal that has little or no constancy of form or intent No attempt to address the task focus		Task focused presentation that considers the individual task objectives but does not justify rationally A consistent proposal that considers the functional layout and data representation without justification against human perception and/or cognition		Task-driven representation focuses on the collective task objectives and reduces the dimensionality of data presented to reflect the task. Justification of approach based on human perception and cognition for functional elements but without consideration of scheme as a whole		Definition and justification of bespoke data representations that define appropriate data-centric displays and features Clear and consistent format and layout with a reasoned and justified perceptive and cognitive feature set throughout A highly objective focused representation that presents all evidence and draws the conclusion for the task objective.	

	0-9%	10-19%	20-29%	30-39%	40-49%	50-59%	60-69%	70-79%	80-89%	90-100%
Level 7 Assessment Scale	Extremely Poor	Very poor	Poor	Inadequate	Unsatisfactory	Satisfactory	Good	Very Good	Excellent	Outstanding
Statistical analysis & Data dashboard: *Critical Review*	None		Personal comment without supporting evidence. No attempt at developing a balanced and reasoned argument based on BCS Code of Conduct and other ethical frameworks General comments, without specific examples and exploration of cases. General discussions on regression, correlation, and hypothesis test results		Functional evidence to support decisions, without discussion of how these were applied or adapted Some attempt at development of balanced argument for the approaches taken based on BCS Code of Conduct and other ethical frameworks, with little/no justification from scientific evidence Supporting research of visual form decisions based on unstructured and validated web-based presentations of data forms Some attempt at development of balanced analysis of regression, correlation, and hypothesis test results		Detailed evidence supports all high-level design ethical principles and presentation decisions that relate general principles to the broad scope of the chosen and selected representations. Discussion on the relationship between task and display narrative with critical evaluation of the features within this. Specific referenced evidence to support design decisions and overall coherence of the visual schema Discussion on the regression, correlation and hypothesis test results		A critical review of all visual representation decisions and factors at both high level conceptual and specific individual representational level A critical review of the matching of the task to the proposed data visualization with consideration of the task and matching of the form presented to the task objectives. A detailed critical review of the BCS Code of Conduct and other ethical frameworks in the context of this data science project Detailed critical analysis of the regression, correlation and hypothesis test results	

Submission Details

The work should be submitted as a single word document with appropriate structure and format, using standard referencing formats throughout. The guideline length for this is 4000 words. You should also submit a **self-assessment of your work**, in the form of a high-lighted marking grid.

Feedback

Feedback in the form of a personalised annotated marking grid and comment sheet will be available within 3 working weeks of the submission date.

The University has strict policies on unfair means. It is your responsibility to ensure that you both understand these and adhere to them in the production of your assignment. Any submitted works with such content identifiable will be penalised in accordance with the University of Salford regulations (http://www.governance.salford.ac.uk/page/academic_handbook).