Assessment Brief and Marking criteria 2022-23

Module title	**Principles of Data Science**
CRN	G340 M0001 / CRN41140
Level	7
Assessment title	**Statistical Analysis and Interactive Dashboard Design**
Weighting within module	This assessment is worth 100% of the overall module mark.
Submission deadline date and time	**09/12/2022 at 4:00 pm.**

Module Leader/Assessment set by
Kaveh Kiani, email: K.kiani@salford.ac.uk
Nathan Topping, email: n.j.topping@salford.ac.uk

How to submit
Your assignment should be submitted through blackboard and should be separated into two formats. First, **a single pdf report** and second, **a zipped file containing your codes and dashboard**. Please check that the report file and zip file are:

1. Your report has been named as "**your name.pdf**".
2. Check the zip file is valid and openable.
3. The zip file should contain material that are clearly labelled and fully working versions of **R codes** and **dashboards** should be included with a clearly written description of each application and its use in a "**Read Me.txt**" file. Your dashboard should be shared as a **.twb** file if you have used Tableau, and **.pbix** if you have used Power BI.

Assessment task details and instructions

Your task is to demonstrate your newly developed knowledge and understanding of data handling, validation, statistical analysis, and visualisation by exploring and presenting data from an extensive and complex data set.
There are two sources of data for this assignment:

1. World Development Indicators (WDI)
2. UNdata

World Development Indicators (WDI) and UNdata are the main World Bank collection of indicators and United Nations Data Bank, compiled from officially recognised international sources. Both sources include national, regional, and global estimates. Both sources include numerous indicators for countries around the world. Also, there are yearly slices of data for the countries that could be construed as time series.

The dataset for this assignment can be accessed from one of these two sources of **your choice:**

https://databank.worldbank.org/source/world-development-indicators

http://data.un.org/Explorer.aspx

Once you have followed the above links, you can download the dataset by selecting the countries and variables you want to work with.

This assessment requires a comprehensive statistical analysis, and a working dashboard prototype to be presented. Your statistical analysis and interactive dashboard design should be fully justified and explained in your report. The assessment also requires you to demonstrate the justified use of techniques for data preparation, validation, analysis and/or modelling and prediction; appropriately referencing research into dashboard composition, layout, function, and form**.** Justification of the approaches taken for statistical analysis and visualisation is expected and outputs should be provided. Your reasoned thinking, research, and critical evaluation of both the problem resolution and your solution also form a substantive part of this work.

Task 1: Interactive Dashboard Design (50% of Total Mark)

For this first task, imagine you are working as a Data Scientist at a Non-Government Organisation involved in social and economic development globally. Part of your role is to use data to communicate these issues to a wider public. As your first assignment at the organization, they have asked you to select some indicators (using the data sources above) which you believe tell a significant story, and to produce a single-screen interactive dashboard to present this data. For example, it could be to compare the trade situation of the least developed countries with developed countries. Your dashboard is to be made publicly available on their website, so you should consider how you can present the data to a general audience who may not have existing expertise in the subject you choose.

The requirements for the proposed dashboard are:

1. Clearly define the objectives of the dashboard based on the dataset you have selected.
2. Based on the objectives, select at least 10 suitable countries of your choice.
3. Produce a single-screen interactive dashboard of at least 10 countries' data.
4. Clear, effective presentation of all factors in a coherent, intuitively comprehensive form, reflecting the objectives you have set for your dashboard.

5. A design 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).

Alongside the dashboard design, you should provide a full report which summarises:

- The objectives you have defined for your dashboard, indicating clearly what your planned solution will communicate to your audience
- The data visualisation principles which have informed your dashboard design, with reference to literature and best practice in data visualisation
- The steps you have taken to pre-process and prepare the data
- An overview of your design with a full justification of the design rationale

For extra credit you should also implement the following advanced features in your dashboard design:

- Use of DAX (if using Power BI) or Calculations (if using Tableau)
- Use of relationships in your data model
- Use of hierarchies, grouping or binning
- Use of in-built Power BI / Tableau forecasting tools

To receive extra credit, these features must be fully documented in your accompanying report.

Task 2: Statistical Analysis (50% of Total Mark)

The requirements for the proposed statistical analysis are:

1. Define research objectives based on the dataset. For instance, to compare the trade situation of the least developed countries with developed countries.
2. Based on the objectives, select at least 10 suitable countries of your choice.
3. Choose a set of indicators according to the objectives 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 R 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 and evaluate the results in the context of your stated objectives.

   4.3. Do regression analysis. Explain why the selected regression techniques are appropriate for the selected variables and defined objectives and show if you've found any similar research in the literature.

   4.4. Do time series analysis. Explain why the selected techniques are better for the defined objectives and show if you've found any similar research in the literature.

   4.5. As a researcher, do a comparative analysis of the main hypothesis testing approaches for your objectives, explain when and why they are used. Then define at least two hypotheses testing related to the objectives 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.

Remarks:

1. You can use a similar datasets and objectives for both tasks. Although. if you prefer you can select different objectives and datasets for each task.
2. You must use R programming language for the entire statistical analysis part (task 2).
3. You can use Tableau or Power BI to develop the dashboard.
4. You can mix different data sets / variables to make your own data set in a meaningful and correct format.

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, various regression models and categorical data analysis, test of hypothesis, and time series analysis).
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 real-world 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 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

5. Technical report writing.
6. Ability to use tools and techniques for statistical analysis.

7. Presenting data in a manner accessible to non-technical stakeholders.
8. 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 between **6000 to 8000 words** (between 30 to 40 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 Dr Kaveh Kiani or Nathan Topping 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 based on university’s reassessment calendar and routines.

Assessment Rubric

Scale	Mark	Rank	Statistical Analysis Description	Data Visualization Description
Outstanding	100 95 92	Distinction	• Aim and objectives have been defined and 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. • Besides general preparation, handling missing data and outlier detection algorithms have been utilized. • In depth descriptive statistical analysis has been provided. • Demonstrating comprehensively the R analysis steps. ➢ Outstanding correlation, regression and time series analysis has been done (2 advance Reg and 2 TS models) ➢ Correct predictions have been made based on 4 models. ➢ Outstanding comparative analysis of the hypothesis testing and more than 2 test of hypothesis has been included. ➢ All the results consist of highly precise and well-explained statements for both technical and non-technical audiences.	• Comprehensive practice in visualisation techniques. • 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. • Comprehensive representation of data using tailored visualisations that extend and refine the basic tool functionality. • Detailed and thorough 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. • A detailed and thorough critical review of the proposed data visualization with consideration of the task and matching of the form presented to the task objectives. • Use of the additional features mentioned in the brief.
Excellent	88 85 82		• Aim and objectives have been defined and 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. • Besides general preparation, handling missing data and outlier detection algorithms have been utilized. • In depth descriptive statistical analysis has been provided. • Demonstrating comprehensively the R analysis steps. o Correct correlation, regression and time series analysis has been done (2 advance Reg and 1 TS models) o Correct predictions have been made based on 3 models. o Excellent comparative analysis of the hypothesis testing and more than 2 test of hypothesis has been included. o All the results consist of precise and well-explained statements for both technical and non-technical audiences.	• Strong practice in visualisation techniques. • 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. • Justification of approach taken, principles adopted, and assumptions made about the representational paradigm. • Comprehensive representation of data using tailored visualisations that extend and refine the basic tool functionality. • Definition and justification of bespoke data representations that define appropriate data-centric displays and features. • Clear format and layout with a justified perceptive and cognitive feature set throughout. • An objective focused representation that presents all evidence and draws the conclusion for the task objective. • A critical review of the proposed data visualization with consideration of the task and matching of the form presented to the task objectives. • Use of the additional features mentioned in the brief.

Very Good	78 75 72		• Aim and objectives have been defined and 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. • Besides general preparation, handling missing data and outlier detection algorithms have been utilized. • In depth descriptive statistical analysis has been provided. • Demonstrating comprehensively the R analysis steps. ❖ Correct correlation, regression and time series analysis has been done (1 advance Reg and 1 TS models) ❖ Correct predictions have been made based on 2 models. ❖ Very good comparative analysis of the hypothesis testing and more than 2 test of hypothesis has been included. ❖ All the results consist of good statements for both technical and non-technical audiences.	• Detailed consideration of comparative analysis presented within the visual representation for a small selection of countries. • Detailed consideration of a common perceptual model and justification of this in a cognitive context. • Refined representation of data using tailored representational forms that extend and refine the basic offering of the packages considered. • Thorough task-driven representation focuses on the overall task objectives and has selected and presented the data to meet these objectives. • A developed justification of approach based on human perception and cognition for functional elements but without consideration of scheme as a whole. • Specific referenced evidence to support design decisions and overall coherence of the visual layout
Good	68 65 62	Merit	• Aim and objectives have been defined and explained. • More than minimum sample size has been used. • Consideration of data preparation. • Besides general preparation, handling missing data and outlier detection algorithms have been utilized. • Good descriptive statistical analysis has been provided. • Demonstrating comprehensively the R analysis steps. • Correct correlation, regression and time series analysis has been done (2 Reg and 1 TS models) • Correct predictions have been made based on 2 models. • Good comparative analysis of the hypothesis testing and 2 test of hypothesis has been included. • All the results consist of good statements.	• Some consideration of comparative analysis presented within the visual representation for a small selection of countries. • Some consideration of a common perceptual model and justification of this in a cognitive context. • Representation of data using tailored representational forms that extend and refine the basic offering of the packages considered. • Task-driven representation focuses on the overall task objectives and has selected and presented the data to meet these objectives. • Justification of approach based on human perception and cognition for functional elements but without consideration of scheme as a whole. • Some referenced evidence to support design decisions and overall coherence of the visual layout
Satisfactory	58 55 52	Pass	• Aim and objectives have been defined. • Minimum sample size has been used. • Minimum data preparation. • Satisfactory descriptive statistical analysis has been provided. • Demonstrating R analysis steps. • Correct correlation, regression and time series analysis has been done (1 Reg and 1 TS models) • 2 test of hypothesis has been included. • Some results have explanations.	• Functional representation of raw data based on standard representations with some modification of the attributes • Indirect representation of data with minimal analysis or pre-preparation. • Some justification of approaches and principles applied. • A consistent but basic report that shows how general principles and approaches have been used to define a coherent presentation. • Task focused presentation that considers the objectives set but does not fully justify • A consistent report that considers the functional layout and data representation without justification against human perception and/or cognition. • Some supporting research of visual form decisions based on unstructured and validated web-based presentations of data forms

Unsatisfactory	48 45 42	Fail	• Aim and objectives have not been defined properly. • Less than minimum sample size has been used. • Minimum data preparation. • Unsatisfactory descriptive statistical analysis has been provided. • Unsatisfactory R analysis steps. • Unsatisfactory correlation, regression and time series analysis has been done (some of the models are wrong) • Requested test of hypothesis has not been included. • Results have not been explained.	• Functional representation of raw data based on standard representations with minimal modification of the attributes • Indirect representation of data with minimal analysis or pre-preparation. • Little justification of approaches and principles applied. • A consistent but basic report that shows how general principles and approaches have been used to define a coherent presentation. • Task focused presentation that considers the objectives set but does not justify rationally. • A basic report that considers the functional layout and data representation without justification against human perception and/or cognition. • Little or no supporting research of visual form decisions based on unstructured and validated web-based presentations of data forms
Inadequate	38 35 32		• Aim and objectives have not been defined. • Less than minimum sample size has been used. • Minimum data preparation. • Inadequate statistical analysis has not been provided. • Inadequate R analysis steps. • Inadequate correlation, regression and time series analysis has been done (Most of the models are wrong). • Requested test of hypothesis has not been included. • Results have not been explained.	• Less than the minimum number of countries have been visualised • Functional data representations using basic forms with minimum modification • Inconstant/uncoherent report that has little or no constancy of form or intent. • Little attempt to address the task focus or set clear objectives for the dashboard
Poor	28 25 22		• Aim and objectives have not been defined. • Less than minimum sample size has been used. • Without data preparation. • Descriptive statistical analysis has not been provided. • Without R analysis steps. • Poor correlation, regression and time series analysis has been done (All the models are wrong). • Requested test of hypothesis has not been included. • Results have not been explained.	• Single country case considered and presented as a worked example. • Functional data representations using basic forms without modification • Inconstant/uncoherent report that has little or no constancy of form or intent. • No attempt to address the task focus or set clear objectives for the dashboard
Very Poor	18 15 12		• Some analysis based on invalid data set has been done. • None of the defined tasks have been done correctly, even according to the minimum expectations.	• Some basic visualisation on invalid data set has been done • Little or no demonstration of understanding of data visualisation principles and relevant tools.
Extremely Poor	8 5 0		• None of the defined tasks have been done correctly, even according to the minimum expectations.	• No demonstration of understanding of data visualisation principles and relevant tools.