Spreadsheet Data Analytics

# Continuous Assessment Project 2021

**Summary**

This project will contribute **50%** towards the final mark for this module.

The working Excel files, presentation and screencast must be submitted through Moodle before midnight on **5th January 2022.** There will then be a short interview on **12th January 2022**.

**Outline**

For the main assignment in Spreadsheet Data Analytics you are required to source a real data set clearly related to an industry problem. You should then solve this problem by working through each of the steps of the Data Analytics Lifecycle we have studied in class using Excel to analyse the data. You can use macros and VBA to automate tasks if you wish. You can also create a customised dashboard to allow the user to investigate the data.

You are encouraged to use the multiple linear regression techniques we investigated in the course. In this regard it may help if the problem you choose includes a predictive element related to a numerical outcome.

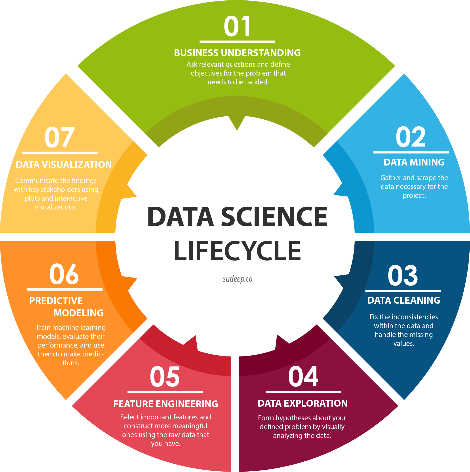
You are required to complete and deliver a report on your work through a Powerpoint presentation targeted at industry colleagues, a screencast that explains the presentation and an Excel workings file of files that detail the data and the steps performed in your analysis.

The initial stage of the project will involve the identification of the data set you wish to use in this project for analysis. This step must be completed by November 24th and agreed with the lecturer. You can of course choose a problem and data set relating to your work but, if you do, you must make sure the data set is anonymised completely. It may also make sense to scale any sensitive data.

For students on the Higher Diploma course, it is recommended that you use the same problem in this assignment and in the corresponding assignment for the Research Methods module.

**Description**

In this project you are required to source a real data set clearly related to a specific problem or problems and to apply each of the steps in the Data Analytics Lifecycle we have studied to solve the problem(s). In solving the problem(s) you should make use of the regression prediction methods we have studied.



Based on the steps in the Data Analytics Lifecycle you should complete each of the following:

**Step 1 - Business Understanding:** Clearly identify and specify what problem(s) the data can be used to solve and explain why this might be important to a particular company and to a broader industry. This problem should involve a prediction element and an associated response variable. However, you will also have associated problems you will need to solve such as the profile of the predictor variables and the relationship between the predictor variables and between the predictor variables and the response variable(s).

Note that the problem set may be expressed as a set of questions or as a set of objectives.

**Step 2 - Data Mining:** Collect the primary data for the project and locate and collect any additional data that might help you solve the problem. Import this data into Excel in a suitable form. Details of the steps and difficulties associated with this stage of the analysis should be included as comments in the notes section of the Excel file you submit. Importantly, you must characterise all the variables within your data set.

**Step 3 - Data Cleaning:** Clean the data using the methods we explored through the course. This activity will naturally require you to deal with missing values and outliers. Again, all activity, issues and decisions should be clearly documented in your file.

**Step 4 - Data Exploration:** You are required to perform Exploratory Data Analysis following the methods we have developed through the course. This will include Univariate, Bivariate and Multivariate Analysis. Through this activity you may well answer some of the questions driving this project including questions relating to profiling the data and understanding the relationship between variables. While the main outputs from this will be summarised in the powerpoint presentation, you should also include detailed comments on your interpretation of each stage of this analysis.

**Step 5 - Feature Engineering:** You are required to identify the important features as they relate to the problems you are trying to solve. This is likely to involve key analysis of the data to address the problems you specified in step 1. This section may also involve the identification of the key variables for completion of the predictive element of the work. You may also need to add additional variables(for example if you wish to convert a categorical variable to a numerical variable) or to construct more meaningful ones.

**Step 6 – Predictive Modelling:** You must use appropriate regression methods to deliver a prediction model for the response variable(s) you are seeking to model. Of course you should examine the quality of the predictions and make conclusions as to the suitability of the model for practical use and its limitations. This step may also involve the fitting of appropriate distributions to a number of the variables to predict future values. You should include statements as to the goodness of fit of chosen distribution.

**Step 7 – Data Visualisation:** Clearly present key findings and generate appropriate visualisations to represent same. You may include dynamic dashboard functionality if you wish. You may also choose to generalise your findings if appropriate.

**Deliverables**

You are required to submit the following through Moodle:

A **Powerpoint presentation** of no more than 12 slides that contains the following:

Slide 1: Introduction and description of data and problem to be solved. This should highlight what companies and industries this problem could relate to. The source of the data should also be given as well as a **link to the screencast for the project**.

Slide 2: Details of Data Mining and Cleaning (Step 2 and 3)

Slides 3 & 4: Details of Exploratory Data Analysis (Step 4) conducted and key results as they relate to the business problems you outlined.

Slides 5: Details of step 5.

Slides 6 - 8: Details of step 6. Details of model(s) developed and results including the quality and the limitations.

Slide 9 - 10: Communication of findings. This is the key slide which should contain visualisations and text to communicate key findings relating to the problem(s) to be solved. It should clearly state the key relationships in the data and the model that should be used to predict the response variable. The structure of these slides should follow the structure of the problems you specified in Step 1.

Slide 11: A final slide that presents what you feel were the most challenging parts of the project.

A **Screencast** of **no more than 10 minutes** where you walk through the entire presentation highlighting any key aspects of the analysis that led to your findings.

An **Excel** dataanalysis file which details the analysis in a structured way and includes a **notes sheet** that logs the sources of the data, details any restructuring and cleaning activity performed and includes a log of when each element of work was completed. You may include version control information in the log if you wish. The Excel file should begin with an **index sheet** which lists the sheets in the workbook and a summary of the content of each. The sheets should be appropriately named based on the steps above.

A **signed declaration form** as below.

**Deadlines**

Sunday 19th December 2021 – Submission of Excel file(s) and Powerpoint file through Moodle with screencast.

Monday 20th December 2021 – Short 8 minute Interview.

**Indicative Marking Scheme**

* Identification and explanation of problem and related companies and industry: 10%
* Data Mining: 5%
* Data Cleaning: 10%
* Exploration of data: 22.5%
* Feature Engineering and Predictive Modelling: 22.5%
* Presentation of findings including Visualisation: 15%
* Overall presentation quality and screencast: 10%
* Interview 5%

**Rubric**

In marking this assessment the following criteria will apply

|  |  |
| --- | --- |
| Grade (%) | Description |
| 80+ | Project demonstrates mastery of subject matter with novel/original work applied to a complex problem.  Data is sourced correctly and has been appropriately structured and analysed.  Analysis of data and model development is accurate, complete, thoroughly explained, and appropriate with clear link to aims. Analysis shows original thinking and implementation beyond what was learned in course. All documentation is well-structured and very well-written. Student has presented with full knowledge of both the problem and the solution including appropriate critical analysis. |
| 70-79 | Project demonstrates thorough understanding of subject matter.  Data is sourced correctly and has been appropriately structured.  Analysis of data and model development is accurate, complete, thoroughly explained, and appropriate with clear link to aims. Functionality is well developed with only minor issues. Document is well-structured, with complete version history. Student has presented with excellent knowledge of both the problem and the solution including appropriate critical analysis. |
| 60-69 | Project demonstrates good understanding of subject matter.  Data is sourced correctly and has been appropriately structured and analysed.  Analysis of data and model development is accurate, substantially complete, well explained, and appropriate with good link to aims. May be missing some appropriate analysis.  Document is reasonably structured, substantially complete and well delivered. Student has presented with reasonably good knowledge of problem and solution but lacks some critical analysis and in depth understanding. |
| 50-59 | Project demonstrates reasonable understanding of subject matter.  Data analysis and model development is accurate, partially complete, reasonably well explained, and appropriate with some link to aims. May be missing some appropriate analysis and links to aims and objectives may not be complete.  Document is adequately structured, mostly complete and reasonably well delivered. Student has presented problem and solution satisfactorily, but with some issues relating to problem and solution. Lacks critical understanding. |
| 40-49 | Project demonstrates partial understanding of subject matter.  Analysis is relatively simple allowing very basic exploration and model development.  Basic but incomplete analysis based entirely on functionality provided in class with significant omissions. Poor link to aims and objectives and weak understanding of logic and findings of the model.  Document missing some key elements. Student has presented poorly. While presentation has some relevant structure, it lacks completeness and coherence. |
| 30-39 | Project demonstrates little understanding of subject matter.  Little evidence of exploration of problem and model development.  Incomplete analysis based entirely on functionality provided in class with weak links to aims. Documentation poor and missing significant elements. Presentation is very confusing and lacks clarity and focus. No critical analysis. |
| 0-29 | Project demonstrates almost no understanding of subject matter.  No serious attempt to address the problem. |

**Additional Notes**

You can of course use analytical techniques you are aware of outside of the module.

You should endeavour to perform as much of the analysis as possible within Excel.

To produce a screencast you can use Zoom. Upload the final screencast with your submission to Moodle.

**Plagiarism**

PLEASE PAY SPECIAL ATTENTION TO THE ISSUE OF PLAGIARISM. The DkIT policies are available at

<https://www.dkit.ie/system/files/academic_integrity_policy_and_procedures.pdf>

in summary, all work submitted by learners for assessment purposes, or for written or oral

publication, must be their own work. Where this is informed by the work of others,

the source must be properly referenced using the accepted norms and formats of the

appropriate academic discipline.

You must sign and submit the declaration below.

**DECLARATION**

**YOU MUST READ, SIGN and upload THIS FILE with your continuous assessment submission.**

|  |  |
| --- | --- |
| **NAME** |  |
| **ID NO.** |  |
| **MODULE** | **Spreadsheet Data Analytics** |
| **ASSIGNMENT** | **CA3 – November 16th 2020** |

I declare that:

* The work submitted for this assessment is entirely my own, and no part of it has been copied from any other person’s words or ideas,
* No part of this assessment has been written or completed for me, it has not been discussed or collaborated on by me with any other person.
* I understand that I am bound by the DkIT Academic Integrity Policy   
  ( <https://www.dkit.ie/system/files/academic_integrity_policy_and_procedures.pdf> ).   
  I understand that I may be penalised if I have violated the policy in any way.