**Part A (Proposal)**

**Weight:** 10%

**Length: 750-1000 words (not including code samples in appendix)**

**Group Assessment**

**Submitted: On Canvas in PDF format by *one group member*. Be sure to follow the naming convention defined at the front of this brief.**

**Task**

In this Assessment item, you will work in a team to produce a project proposal for your data analysis project. You will do this by following this sequence of steps:

1. Define a broad research area of interest (e.g. public health, climate change, demographic change, finance etc.) and form groups accordingly.
2. Work to define a set of well specified research questions for your broad area of interest. (Note: at least some of these should be actionable – see Part B).
3. Look for a range of datasets that might help to answer their questions.
4. Refine at least one research questions so that it can be answered by a **Regression Model**. (Note: you can ask other questions too, but you must produce at least one regression model in Assessment 2B.)
5. Write a proposal that summarises the following:
   1. the rationale and stakeholders for the project,
   2. the research questions,
   3. the range of datasets examined as well as those chosen for the analysis (include details about how you merged the different datasets and an assessment on whether the granularity of the data sources is sufficient to answer your research questions),
   4. the regression modelling techniques to be employed and,
   5. any issues that you anticipate might arise in carrying out the project.
   6. Include an Appendix that contains code samples demonstrating the data acquisition and merger processes that you have used to date.

**Assessment Criteria: Part A**

|  |  |  |  |
| --- | --- | --- | --- |
| **SLO** | **CILO** | **Assessment Criteria** | **Weight** |
| 2 | 1.4 | Clarity in articulating the research questions along with a well defined proposal for making the invisible visible for a specified set of stakeholders. | 30% |
| 4 | 1.2 | Level of expertise using key R functionality demonstrated in the process of data acquisition, and creativity in solving the problem of finding and merging datasets that can answer the research questions. | 50% |
| 3 | 4.2 | Eloquence and robustness of the argument used to justify the proposal. | 20% |
| **Sub Total** | | | **100** |
| **Total (10%)** | | | **/10** |

# Introduction and research questions

Intro to topic and stakeholders

The Data Geeks have identified four research questions that may be explored to provide the business stakeholders with insights into the factors effecting crop yield production:

1. How have weather events effected crop yield production in Australia?
2. What geographical factors influence the yield of crops grown in Australia?
3. How will climate change factors affect crop yield production?
4. How do different farming practices around the world effect crop yield production?

Based on preliminary research and exploration of the data available Question 2 has been selected as the topic that the Data Geeks will aim to answer using regression modelling techniques.

# Background

List of factors that affect crop yield – Marco

# Data collected

Table of data that lists variables, source – Rato, All to add the sources they have collected

The Data Geeks have collected and analysed the following sets of data which are grouped into 7 main area:

* 1. Crop yield in Australia regions
     1. Cereal crop yield vs nitrogen fertiliser application - by country, year (1961 to 2018)
     2. Average of farm yields (tonnes / hectare) in Australia
        + For crops and livestock
        + From 1979-2019 (actual) and 2020-2021 (forecast)
     3. Summary of wheat in Australia
        + From 1973-2018
        + Area (hectares)
        + Yield (t/ha)
        + By state
     4. Rainfed wheat
        + By station (geo-spatial
        + By climate zone
        + By country
     5. Crop production and area in Australia
        + From 1998 to 2018
        + Winter and Summer crops
        + By state
     6. Trading of cereals 1961 to 2018
     7. Crop yield in US from 1993 to 2019
     8. Crop value in US from 1998 to 2017
  2. Land use
     1. for agricultural production, crops, grazing, and forestry (ABS, 46270DO002\_201617 Land Management and Farming in Australia, 2016–17)
     2. from CSIRO Data Access Portal (<https://data.csiro.au/dap>)
  3. Fertiliser used in regions for different types of land use
     1. Fertiliser by state in Australia for 2016-17
  4. Soil data
     1. as fetched from SLGA website as grid data (<https://www.clw.csiro.au/aclep/soilandlandscapegrid>)
     2. Soil attributes
        + Available Water Capacity
        + Bulk Density (Fine Earth)
        + Bulk Density (Whole Earth)
        + Cation Exchange Capacity
        + Cation Exchange Capacity (Effective)
        + Clay
        + Coarse Fragments
        + Depth of Regolith
        + Depth of Soil
        + Electrical Conductivity
        + Organic Carbon
        + pH CaCl2
        + pH Water
        + Sand
        + Silt
        + Total Nitrogen
        + Total Phosphorus
  5. Climate data and climate change projections
     1. Temperature – min, max for 1961-90 and projection into 2030, 2050, 2070 and 2090; in raster files
     2. Rainfall – min, max for 1961-90 and projection into 2030, 2050, 2070 and 2090; in raster files
     3. Solar radiation – min, max for 1961-90 and projection into 2030, 2050, 2070 and 2090; in raster files
  6. Geo-spatial data
     + - Yield area – for 2015-16, in CSV
       - Land use – for 2018, in shape files
       - SA2 – for 2016 Australia, in shape files
  7. Labour force
     1. employed and expenditure by type of work, industry for 2015-16
     2. employed by country by year

# Proposed methodology

The Data Geeks will developed a Generalised Linear Model (GLM) with the aim of answering the question; What geographical factors influence the yield of crops grown in Australia? A multivariable regression model will be developed to predict crop yields for specific regions of Australia. The main groups of predictor variables that will be used are climate data, soil data and fertiliser use.

The datasets will be combined by analysing into a consistent format by sampling the gridded climate and soil data

# Project risks

Any potential issues or difficulties

Table List of project risks and

|  |  |  |
| --- | --- | --- |
| Risk | Description | Mitigation/Alternative methodology |
| Geospatial data is not granular enough |  |  |
| Mismatching of geospatial regions |  |  |
| Crop yield data and landuse inaccuracies/omissions |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# Appendix

R code for data acquisition and joining

# Notes

## Response Variables

Yield of Wheat of State = wheat production (‘000 tonnes) / agricultural land area (hectares)

## Predictor Variables

* + 1. Rainfall[[1]](#footnote-1) – April (early sowing season)
    2. Rainfall – May (normal sowing season)
    3. Rainfall – June (normal sowing season)
    4. Rainfall – July (late sowing season)
    5. Rainfall – August (growing stage)
    6. Rainfall – September (flowering stage)
    7. Rainfall – October (harvesting stage)
    8. Rainfall – November (harvesting stage)
    9. Temperature – April (early sowing season)
    10. Temperature – May (normal sowing season)
    11. Temperature – June (normal sowing season)
    12. Temperature – July (late sowing season)
    13. Temperature – August (growing stage)
    14. Temperature – September (flowering stage)
    15. Temperature – October (harvesting stage)
    16. Temperature – November (harvesting stage)
    17. Solar Radiation – August (growing stage)
    18. Solar Radiation – September (flowering stage)
    19. Solar Radiation – October (harvesting stage)
    20. Solar Radiation – November (harvesting stage)
    21. Fertilisers used
    22. Soil condition
    23. Labour force employed

1. GrowNote-Wheat-South-03-Planting.pdf [↑](#footnote-ref-1)