**Assessment Task 2: Data analysis project**

**Weight: 50% (see below for breakdown to different parts and due dates)**

**Rationale**

As a data scientist one of your core duties will be to work with a team of people to analyse complex datasets, and to report back the results of that analysis to stakeholders from a variety of backgrounds, who often have different needs and capabilities. This assessment task will give you a chance to experience the complexity that can often arise in this situation.

You **must keep a timestamped reflective journal** throughout your groupwork project. You should do this using the GoingOK tool which is available at: <http://goingok.org/>

**Part A (Proposal)**

**Weight:** 10%

**Length: 750-1000 words (not including code samples which should be included in an appendix)**

**Group Assessment**

**Due date: 11:59pm Monday 29th April**

**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 **Predictive Regression Model**. (Note: you can ask other questions too, but you must produce at least one regression model in Assessment 2 Part B.)
5. Write a proposal that summarises the following:
   1. the rationale 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 might be merging different datasets here),
   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 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 | The clarity of the research questions articulated, with a clear intention to make the invisible visible for a well defined set of stakeholders. | 30% |
| 4 | 1.2 | The process of data acquisition demonstrates mastery of key R functionality (e.g. interfacing with APIs) and a creative attempt to solve the problem of finding data that can answer the research questions. | 50% |
| 3 | 4.2 | An eloquent and well justified proposal. | 20% |
| **Sub Total** | | | **100** |
| **Total (10%)** | | | **/10** |

**Submission Requirements:**

* All assignments need to be submitted via Canvas unless otherwise instructed
* Submissions must include a title page with subject named, student name(s) and IDs, date of submission and the title of the assessment.
* Please use the following file naming format for each submission:
* For individual assignments: StudentName\_AssignmentName\_Date.
* For group assignments: TeamName\_AssignmentName\_Date
* Please use embedded objects instead of linked objects for content sourced externally

|  |  |
| --- | --- |
| Description: 18897 CIC A4 Portrait WordTemp_cropped.jpg | **ASSIGNMENT COVER SHEET** |
|  | |
|  | |
| |  |  | | --- | --- | | **SUBJECT NUMBER & NAME** | 36103 Statistical Thinking for Data Science | | **STUDENT NAMES & IDs**  **(SURNAME, FIRST NAME, STUDENT ID)** | Arunachalam, Abishek (SID: 13001262)  Jiang, Benjamin (SID: 12875314)  Pelayre, Anne Gorge (SID: 13102191)  Raghavan, Saminathan (SID:13075597)  Ramal, Miguel (SID: 00060259) | | **TEAM NAME** | SKEPTICS | | **STUDENT EMAIL** | [abishek.arunachalam@student.uts.edu.au](mailto:abishek.arunachalam@student.uts.edu.au), [benjamin.jiang@student.uts.edu.au](mailto:benjamin.jiang@student.uts.edu.au), [annegorge.pelayre@student.uts.edu.au](mailto:annegorge.pelayre@student.uts.edu.au), [saminathan.raghavan@student.uts.edu.au](mailto:saminathan.raghavan@student.uts.edu.au), [miguel.ramal@student.uts.edu.au](mailto:miguel.ramal@student.uts.edu.au) | | **DUE DATE** | 30 April 2018 | | **ASSESSMENT ITEM NUMBER/TITLE** | AT2 Data analysis project, Part A: Project Plan | | * I/We confirm that the work submitted conforms with the university’s guidelines on academic integrity.   *Refer to the UTS policy on ‘Advice to Students on Good Academic Practice’*: <http://www.gsu.uts.edu.au/policies/academicpractice.html>   * I/We am aware of the penalties for plagiarism. This assignment is my/our own work and I/we have not handed in this assignment (either part or completely) for assessment in another subject. * If this assignment is submitted after the due date I/we understand that it will incur a penalty for lateness unless I have previously had an extension of time approved and have attached the written confirmation of this extension.   Please provide details of extensions granted here if applicable \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **Signature of Team:** \_\_\_\_\_\_Skeptics\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ **­­Date:**  30 / 04 / 2018  If submitted electronically tick here to indicate you agree with the above X | | | |

**Rationale for project:**

Gross Domestic Product (GDP) is the most popular method for measuring an economy’s output, a measure of the size of a country’s economy. For this assignment our team’s decision to analyse and predict the GDP of Australia from its economic indicators is an attempt to provide insights and understanding how GDP gets calculated, the data currently available to help analyse our country’s economy as well as provide some understanding on the factors to consider when analysing macro-economic data.

Our team’s quest will become a reflection on considerations, for data analysts without formal qualifications in economics, to take when dealing with projects requiring analysis of macro-economic data.

GDP as defined in the Australian System of National Accounts is “the total market value of good and services produced in Australia after deducting the costs of goods and services used up (intermediate consumption) in the process of production, but before deducting allowances for the consumption of fixed capital (depreciation).” (Australian Bureau of Statistics 2012)

**Research questions:**

* Can GDP for Australia be forecasted for the next few years?
* What factors may help or deter the forecasting method ?
* Possible:
  + Are there other ways to measure GDP
    - According to this URL from ABS there are 3 methods: <http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/1301.0~2012~Main%20Features~Defining%20and%20measuring%20GDP~221>

**Range of datasets examined and chosen for analysis**

(include details about how you might be merging different datasets here),

After analysing some different sources of datasets for our project (\*list some of these\*), we decided to use as main source of economic government data, a selection of datasets obtained from the Australian Bureau of Statistics (ABS). The ABS site provides a free tool: ABS.Stat that allows for a web browser interface to display, query and download data on multiple themes such as Economy, Health, Industry, Labour, People, Census and other snapshots of Australia. The site also allows for web services interface enabling for data to be accessible for research via APIs using Statistical Data and Metadata Standard (SDMX) format which was chosen as our preferred mechanism to obtain data from ABS. See <http://stat.data.abs.gov.au/> for further information.

Our chosen datasets for analysis from ABS.Stat come from the ‘Economy’ theme: Business Indicators and Price Indexes and Inflation (\*still to add more details here\*)

* **Business Indicators SDMX**

<http://stat.data.abs.gov.au/restsdmx/sdmx.ashx/GetData/QBIS/10+50+90+110.TOTAL.0.99.10+20+30.Q/all?startTime=1985-Q1&endTime=2017-Q4>

For unemployment data we source our dataset from:

* **Unemployment**

https://data.oecd.org/unemp/unemployment-rate.htm

**Regressions Modelling techniques**

**Issues that may arise during project**

**Appendix**

(contains code demonstrating the data acquisition and merger processes that you have used to date)

Insert chosen snippets of code from the 3 main R scripts:

Business\_Indicators.R

Price\_Indexes.R

Unemployment.R

And a sample of data transformation from:

Goods and services - monthly data into quartiles.R

**References**

Australian Bureau of Statistics 2012, *Defining and measuring GDP*, viewed 28 April 2018, <http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/1301.0~2012~Main%20Features~Defining%20and%20measuring%20GDP~221>.

Aspden, C 1990, ["Which is the best short-term measure of Gross Domestic Product?"](http://www.ausstats.abs.gov.au/ausstats/free.nsf/0/A883157CBF2E2E21CA2574DC0016C234/$File/52060_1990_JUN.pdf) originally published in [Australian National Accounts: National Income and Expenditure, June Quarter 1990](http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/5206.0main+features1jun+1990), (5206.0), pp. 57–65