**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 represents the total dollar value of all goods and services a country produced over a specific time period, often referred to as the size of the country’s economy. As one of the most widely used economic indicators, GDP is used to gauge the health of country’s economy (Investopedia 2018). Given the importance of having a healthy economy to the wellbeing of a country’s citizens, our team considered whether it was possible to predict future GDP of Australia using other historical economic and non-economic information (further discussed in datasets section below).

Our team viewed choosing this topic as a learning opportunity, to better understand how we as individuals and a as a community contribute to the economy. As a team of data analysts without formal qualifications in economics, it will decode economic jargon and provide us with insights on what are the important factors to look at when trying to analyse a country’s economy.

**Research questions:**

As there are two types of GDP that economists use to measure a country’s economy, our regression model will disregard real GDP (economic output adjusted for the effects of inflation) solely focus on predicting nominal GDP (a country’s economic output without an inflation adjustment).

The main research question our team came up with was “can nominal GDP be accurately predicted from other economic and non-economic information”.

The secondary research question supplementing this was which economic or non-economic factors are most important for influencing nominal GDP.

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

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.

Whilst measuring GDP can be complicated, at its most basic mathematically, the calculation can be done in one of three ways: either by adding up what everyone earned in a year (income approach) or by adding up what everyone spent (expenditure method), or by how much everyone produced (production approach). While each approach should, conceptually, deliver the same estimate of GDP, if the three measures are compiled independently using different data sources, then different estimates of GDP result (ABS 2012). To combat this issue, the estimates in the GDP dataset we used had been pre-balanced by the ABS between the three approaches.

As it did not matter then which method we choose as long as we were consistent in our logic, we choose the expenditure method as it had the most readily available information for calculating GDP = Consumption + Investment + Government spending + Net Exports.

In doing so, when choosing our data sets, we were also careful to not choose datasets that were components of each of the methods but rather indicators for the components. For example, we used Consumer Price Index and Retail Trade Turnover as an indicator of the level of Consumption in the economy.

|  |  |  |
| --- | --- | --- |
| Expenditure Approach | Indicator | Link |
| GDP | GDP | http://www.abs.gov.au/ausstats/abs@.nsf/mf/5206.0 |
| Consumption | Consumer Price Index | http://www.abs.gov.au/ausstats/abs@.nsf/mf/6401.0 |
|  | Retail Trade Turnover | http://www.abs.gov.au/ausstats/abs@.nsf/mf/8501.0 |
| Investment | Bank Accepted Bills/Negotiable Certificates of Deposit-3 months | https://www.rba.gov.au/statistics/tables/ |
|  | Employed total persons | http://www.abs.gov.au/ausstats/abs@.nsf/mf/6202.0 |
| Government | Unemployed total persons | http://www.abs.gov.au/ausstats/abs@.nsf/mf/6202.0 |
|  |  |  |
| Net Exports | Trade Weighted Index | https://www.rba.gov.au/statistics/tables/ |
|  | Balance on Goods and Services | http://www.abs.gov.au/ausstats/abs@.nsf/mf/5368.0 |

**Regressions Modelling techniques**

**Issues that may arise during project**

**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>.

Investopedia 2018, *What is GDP and why is it so important to economists and investors?*, viewed 28 April 2018, < https://www.investopedia.com/ask/answers/199.asp>.

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

**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