

# ALT 2 Project Brief

## Spring 2019

### Dates of Note

- ALT 2 Specification Publish Date : 27th January 2019
- ALT 2 Project Due : 8th March 2019



### Theme

Your team is required to perform a data analysis project. This project should put forward a hypothesis that should be then tested using data analysis tools and a conclusion reached. The topic of your analysis can be anything of your choosing. Your project development should follow the **Data Science Arc** process and check off each of the listed **Learning Outcomes**.

### Cross Curricular Opportunities

This task has potential crossovers in many subjects. Mathematics will feature heavily - from statistics in the calculation of measures of centrality (mean, mode, medium) and spread (standard deviation, interquartile range) to inferential statistics in your hypothesis testing and conclusions. Depending on your choice of topic you may also see crossovers in other subjects such as history, geography, politics and society, CSPE, etc.

**Marking Outline** You will be marked collectively as a two person team on this project.

- **Planning Document - 10%**
  - Ask : Development of and background to your hypothesis / question.
  - Preparation and planning of project : What data sources will you need? What skills will you need when it comes to development?
  - Get Data : gather, find or create appropriate data set(s)
- **Project Development & Coding - 60%**
  - Cleaning : of the data and creation of cleaned data files.
  - Analysis : of the data using prebuilt / custom tools including measurements of spread and centrality, etc.
  - Comparison of data sets if necessary.
  - Visualisation : of analysed data to illustrate findings.
  - Program comments and code efficiency
- **Final Report & Presentation 15%**
- **Above and Beyond Functionality 15%**
  - E.g. analysis of multiple datasets, manual creation/recording of raw data, integration of addition technology, advanced data visualisations beyond typical bar/line/pie charts.

## Submission of Project

You are required to submit your project and all documentation on or before the 8th of March 2019 via the appropriate assignment channel on Repl.it. Documentation should be shared via the appropriate ALT folder in Google Drive and must include complete Code, original unedited data sets and cleaned data files.

## Requirements

You are required to study a chosen data set and make an hypothesis related to the topic. You are then required to code a python program(s) that imports and analyses the data set allowing you to draw a conclusion in regards your original hypothesis. You are required to produce planning documentation as well as a final report as detailed below.

## Planning Document

- Project Title.
- Overview of the topic and your interest in it. (min 1 paragraph)
- Your hypothesis and explanation. (min 1 paragraph).
- Origin and background of your chosen data set.
- Team Members & Assigned Roles.
- All rough work / sketching / brainstorming notes to be included.

## The Project

We will be following the Microsoft Data Arc process as a guide for our projects as follows.



- Create an hypothesis and write your idea as a question. Ideally this hypothesis should tie together at least two different data points. What data would you need to source to test your hypothesis? Some examples of hypotheses;
  - Is there a correlation between the production cost of a movie and it's takings at the box office?
  - Do crime statistics go up during recession times and go down during prosperous years?

- Gather or generate an appropriate data source / data set to investigate your hypothesis.
  - The data set should have a minimum of 300 data points
  - This data can be obtained from a third party or collected directly through observation / monitoring /sensing.
- Clean your dataset
  - Remove irrelevant data columns
  - Replace inappropriate data points such as NULL, N/A, etc.
  - Convert to appropriate data types.
  - Save as a new 'cleaned' file.
- Analyse the data
  - Identify patterns and trends.
  - Calculate means of centrality or spread as appropriate.
  - Evidence that proves/disproves your original hypothesis.
- Visualise your data
  - Represent your data visually using an appropriate coding library such as matplotlib or other medium.
  - If your data is location related, you can plot points on a map, etc.
- Review your hypothesis
  - Did the evidence support your hypothesis or show otherwise?
  - Summarise your findings.

## **Project Final Report**

- Overview of your project results and the final conclusion on your findings and how they affected your original hypothesis.
- Any opinions or evidence of causality.
- Complete project code.
- All generated visualisations.
- Outline of team member roles in production of project.
- Reflection on planning versus implementation.
  - What parts of the project worked out as expected?
  - What parts of the project did not work out as expected?
  - What were the biggest challenges of the project?
  - What unexpected questions arose during your project?
- Class Presentation (6 slide maximum)
  - Overview of your chosen topic and why you selected it.
  - Overview of the data set(s) you used.
  - Your cleaning process.
  - Visualisations of your data.
  - Conclusions and findings.
  - Any additional functionality.

## Learning Outcomes

### S1 : Computational Thinking

- 1.3 Solve problems by deconstructing them into smaller units using a systematic approach in an iterative fashion.
- 1.4 Solve problems using skills of logic.
- 1.6 Explain the operation of a variety of algorithms.
- 1.7 Develop algorithms to implement chosen solutions.
- 1.9 Use modelling and simulation in relevant situations.

### S1 : Designing and Developing

- 1.19 Identify features of both staged and iterative design and development process.
- 1.20 Collaborate and assign roles and responsibilities within a team to tackle a computing task
- 1.21 Read, write, test and modify computer programs.
- 1.23 reflect and communicate on the design and development process.

### S2 : Abstraction

- Use a range of methods for identifying patterns and abstract common features.
- Implement modular design to develop hardware or software modules that perform a specific task.

### S2 : Algorithms

- Use pseudo code to outline the functionality of an algorithm.
- Construct algorithms using appropriate sequences, selections/conditionals, loops and operators to solve a range of a problems, to fulfil a specific requirement.
- Implement algorithms using a programming language to solve a range of problems.
- Assemble existing algorithms or create new ones that use functions (including recursive), procedures and modules.

### S2: Data

- Use data types that are common to procedural high-level languages.
- Collect, store and sort both continuous and discrete data.

### S2 : Evaluation and Testing

- Test solutions and decisions to determine their short-term and long-term outcomes.
- Identify and fix/debug warnings and errors in computer code and modify as required.
- Critically reflect on and identify limitations in completed code and suggest possible improvements.

### S3 : ALT 2 Analytics

- Develop algorithms that can find the frequency, mean, median and mode of a data set.
- Structure and transform raw data to prepare it for analysis.
- Represent data to effectively communicate in a graphical form.
- Use algorithms to analyse and interpret data in a way that informs decision-making.