# *Project 1*

| | |
|---|---|
| ***Deadline:*** | Hand in by midnight Sunday, 22th of March 2020. |
| ***Evaluation:*** | 15% of your final course grade. |
| ***Late Submission:*** | See Course Guide. |
| ***Work*** | This assignment is to be done **individually**. |
| ***Purpose:*** | Gain experience in perform data wrangling, data visualization and introductory data analysis using Python with suitable libraries. Begin developing skills in formulating a problem from data in a given domain, asking questions of the data, extracting insights and presenting findings from real-world problems. Learning outcomes 1 and 2 from the course outline. |

***Project outline:***

This project requires that you perform introductory exploratory data analysis (EDA) on a real-world problem. Preferably the dataset will be current, topical and original. You are required to formulate a set of questions which you want answered from the data, which will provide you with findings that you must present as a coherent story using a Jupyter Notebook.

Your tasks are:

1. Find or put together a real-world dataset(s) describing some phenomenon that is of interest to you.
2. Once you have chosen your dataset or problem domain, record its name and description on the Google Drive document that has been linked on Stream. Each student will work on a separate dataset and this is to ensure this occurs.
3. Study your dataset and the domain from which it originates. Try to understand the domain as much as you can.
4. Check the data integrity by looking for means, medians, outliers, missing data.
5. Begin to formulate questions that you would like answered from the data and patterns/associations you would like to investigate. Ensure that your work/research questions are original.
6. Comment and discuss each answer/graph you generate and as you move from one insight to the next.
7. Create a Jupyter Notebook, write an abstract, introduction, research questions and the body with all your analysis with a discussion, and finally a conclusion with your key findings. This final notebook must have all the scripts in it that will make your findings repeatable when being marked.
8. Structure your notebook into clear sections and thus ensure readability.
9. Submit both your final notebook as well as the dataset(s) (if they are not too large).

Use data wrangling when appropriate: to transform data into different formats, create new columns as derivatives from others, fill in missing values, transpose and aggregate etc.

Experiment with a variety of initial data analysis and EDA techniques: find means, medians, standard deviations, frequency distributions, min/max values and range, distribution types, pivot tables, group by operations, regression and correlation etc.

Utilize a wide spectrum of visualization graphs/tools: histograms, bar /box graphs, scatter plots, pie/line graphs, static/dynamic.

You may install and use any additional Python packages you wish that will help you with this project. When submitting your project, include a README file that specifies what additional python packages you have installed. Use whatever Python version or distribution that you like.

***Requirements:***

Your notebook must have an introduction to your research (containing the purpose of your research, your key research questions, a brief summary of your with findings), EDA together with explanations and discussions of the findings and a conclusion.

All the figures and calculations must be repeatable when your project is being marked, and all your figures must be explained/interpreted. The figures should have titles and labels so that they are readable and meaningful. Make sure that all the figures and tables are interpreted with sufficient analysis for the reader.

Ensure that your submitted notebook has all the output embedded in it in case there are problems with your code/imported modules and I am not able to execute the code.

Do not 'dump' large amounts of data into the output cells of the notebook as this disrupts the flow. This means that you should not output more than 5 rows of any DataFrame.

**DO NOT:**
Please do not use datasets from Kaggle for this project, or commonly explored datasets such as Titanic. You are asked to produce an original piece of work which requires of you to go beyond simply replicating research which others have done.

***Marking criteria:***

Marks will be awarded for different components of the report using the following rubric:

| Component | Marks | Requirements and expectations |
|---|---|---|
| Data Wrangling | 15 | Thoroughness in data cleaning/preparation/transformations using various techniques. Clean and readable code with consistent naming conventions. Appropriate handling of missing data where necessary. Removal of code duplication and use of functions. |
| EDA/Visualisation | 25 | Variety of exploratory research and inquiry into different aspects of the dataset. Use of broad and appropriate range of visualisations and their effective communication. Judicious use of colour in figures. Checking of data integrity. Identification of unusual or problematic aspects of the data. |
| Data Analysis | 35 | Quality of the questions being asked. The degree to which the research has gone beyond just the summary statistics. Diversity of techniques used to answer the research questions and to present the answers. Depth of discussion and interpretations of findings. |
| Originality | 15 | Originality of the dataset, research questions and the code. The degree to which the problem domain is topical. |
| Presentation | 10 | Structure, readability and flow of the report. The ability to tell a story and persuade the reader as to the significance of the findings. |

**Hand-in**: Make sure that the notebook you submit has all the outputs embedded. Also, export your notebook into HTML. Z**ip**-up all your **notebooks (.ipynb and .html)** and **dataset(s)** you have chosen, as well as any other **.py files** you might have written, into a single file and submit through Stream. Do not email your submission to the lecturer unless there are problems with the submission site.

**If you have any questions or concerns about this assignment, please ask the lecturer sooner rather than closer to the submission deadline.**