Project 1

Deadline:	Hand in by midnight March 27 2016
Evaluation:	5% of your final course grade.
Late Submission*:*	10% deduction from your final mark per day late.
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 from a real-world dataset. Learning outcomes 1, 3 and 4 from the course outline.

Project outline:

This project requires that you perform data cleaning and exploratory data analysis (EDA) on a dataset from a real-world domain. You are to present your work in the Jupyter Notebook which is to have the structure of a report, together with all the Python scripts embedded in it, and descriptions of the steps you are taking in your analysis and the data cleaning process. Discuss assumptions you make and questions that arise as you work with the data.

Transform data into different formats where necessary, create new columns as derivatives from others, fill in missing values where you think it is meaningful, transpose and aggregate, report on percentages of missing values for each feature etc.

Utilise a variety of initial data analysis and EDA techniques. Make use of a wide spectrum of visualization graphs/tools.

You may install and use any additional Python packages you wish that will help you with this project.

The dataset that you will be working on has been provided to us by an industry partner, who has requested to remain confidential. The dataset has been granted access to you under our non-disclosure agreement with our partner, with the understanding that you will not distribute it in any form.

The dataset name is RURAL_LS_SAMPLE.csv. It has 65,000 anonymised records of individuals' responses to various surveys in rural NZ. The feature labels have been made as self-descriptive as possible. In cases where they are not, attempt to understand the context through the feature values and record your assumptions.

Below are some clues to the survey feature names that have been revealed to us:

1. features with '_M' in them indicate that the response refers to the person who was filling out the survey, while '_P' in the feature name refers to their partner.
2. some features have '_A' in them, indicating agree, '_SD' strongly agree, '_D' disagree, '_NO' no opinion or '_SA' strongly agree.
3. INT_ indicated an individuals' interests in lifestyle
4. OCC_ indicates an individuals' occupation

Marking criteria:

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

Component	**Marks**	**Requirements and expectations**
Data Wrangling	40	Thoroughness in data cleaning, use of user-defined functions.
EDA/Visualisation	30	Variety of exploratory research and inquiry into different aspects of the dataset, use of broad and appropriate range of visualisations and their effective communication.
Data Analysis	20	Quality of the questions being asked, diversity of techniques used to answer and present them.
Presentation	10	Structure of the report.

Hand-in: Submit your notebook file via Stream assignment submission link.