Project 1

Deadline:	Hand in by midnight March 24 2019
Evaluation:	10% of your final course grade.
Late Submission:	Refer to the 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 from a real-world dataset. Learning outcomes 1, 3 and 4 from the course outline.

Please note that all data manipulation must be written in python code in the Jupyter Notebook environment. No marks will be awarded for any data wrangling that is completed in excel.

Please do not copy the work of others – there are many ways to solve the problems below, and we expect that no two answers will produce the same code. Copying the work of others (even if object/variable names are changed) will be considered plagiarism.

Question 1: NZ Health survey data - Wrangling, reshaping, functions and plotting (35 marks)

This question relates to the "health_survey.csv" dataset, that you can download from the Stream site. This dataset is from the NZ Health survey. You can learn more about this dataset and what the column names and labels represent here:

https://www.health.govt.nz/nz-health-statistics/national-collections-and-surveys/surveys/new-zealand-health-survey

a) **Importing data:**

- Read in the health data and save to a dataframe object. There is an encoding argument that can be set when using read_csv. You may need to set this to 'latin' to avoid the "utf-8' codec can't decode…" error (see the Pandas documentation for further information).
- Remove the first unnamed column and the seven 'p.value' columns.
- Change **all** 'percent' column names to their associated 'Year' values (e.g. the name for 'percent.16' changes to '2016') and change the column name of 'short.description' to 'description'.

(3 marks)

b) **Filtering:**

- Display the unique labels in the 'description' column so that you can inspect them. Comment out the command that does that and run the code box again once you have finished this question, as the list is large and will clutter your notebook.
- Save a new dataframe object (with an appropriate name) into a new memory location, that contains all the rows that meet **all** the following criteria:
    - That match six of the 'description' labels **of your choosing**. For instance, if you were interested in knowing about 'Physically active', 'Anxiety disorder', 'Daily smokers', 'Diabetes', 'Healthy weight', and 'Self-rated health - very good', then your dataframe would contain only rows that matched these 'description' labels,
    - That also match the 'Total' label in the Group column
    - That also match the 'adult' label in the population column

Note: this is not an 'either/or' filter. All rows in your dataframe must meet all the above conditions. Only a maximum of half marks will be awarded if you choose the exact same 'description' labels as given in the example above. The result should be a six-row dataframe and look like the result pictured below, but with your chosen 'description' labels.

(4 marks)

	population	description	group	2016	2015	2014	2013	2012	2011	2006	2007
0	adults	Physically active	Total	50.2	47.7	50.7	52.1	51.7	54.4	52.0	NaN
112	adults	Anxiety disorder	Total	10.3	9.5	7.8	8.4	6.4	6.1	4.3	NaN
292	adults	Daily smokers	Total	13.8	14.2	15.0	15.7	15.6	16.3	18.3	NaN
386	adults	Diabetes	Total	5.6	5.8	6.1	5.4	5.8	5.5	5.1	NaN
887	adults	Self-rated health - very good	Total	40.9	40.6	40.0	44.0	38.8	37.3	41.0	NaN
1263	adults	Healthy weight	Total	31.9	32.0	33.1	33.4	33.7	34.3	36.2	NaN

c) **Wrangling and reshaping:**

- Now wrangle and reshape the dataframe that you produced in c) until you end up with the result pictured below. Your column names and cell values will be different as you will have chosen different 'description' labels. Please note that the index is 'Year' and it is sorted in ascending year order (you can find how to sort indexes in the documentation). No marks will be awarded for solutions that involve hard-coding cell values.

  (12 marks)

description	Physically active	Anxiety disorder	Daily smokers	Diabetes	Self-rated health - very good	Healthy weight
Year						
2006	52.0	4.3	18.3	5.1	41.0	36.2
2007	NaN	NaN	NaN	NaN	NaN	NaN
2011	54.4	6.1	16.3	5.5	37.3	34.3
2012	51.7	6.4	15.6	5.8	38.8	33.7
2013	52.1	8.4	15.7	5.4	44.0	33.4
2014	50.7	7.8	15.0	6.1	40.0	33.1
2015	47.7	9.5	14.2	5.8	40.6	32.0
2016	50.2	10.3	13.8	5.6	40.9	31.9

d) **Creating a function:**

- Write a function that will perform the wrangling from b and c for any combination of population, group and any list of description labels of any length, e.g. population = 'adults', group = 'Women', description = ['Past-year drinkers', 'Amphetamine use',…].
- Test the function with some different values for your arguments and show that it returns dataframes like you produced in c) (Be sure that your combination of 'group' and 'population' and 'description's makes sense when testing, otherwise you will return empty dataframes – for instance, if you are choosing 'children' as your population, it would not make sense to choose 'Men' as your group) .

  (6 marks)

e) **Plotting:**

- Create two plots of your choosing by using one or more of the dataframes that you have created. Ensure they are appropriate for encoding the data you are displaying. Also ensure they are correctly labelled, with appropriate axis names and tick labels, and a title. We should be able to understand what the plots are showing us without having to read the code.

- In markdown boxes, write some brief insights.

(10 marks)

Question 2: NZ Top 100 baby names time series - Grouping and plotting (20 marks)

This question is to be answered using the "baby_names.csv" dataset available on the Stream site. This time-series dataset has been lovingly pre-wrangled for you. The data is the top 100 baby names by gender, sourced from the Department of Internal Affairs: https://smartstart.services.govt.nz/assets/files/Top-100-girls-and-boys-names-since-1954.xlsx.

- Import this data, and create 3 plots from it, with appropriate labelling, titles, etc. At least one of your plots should use data that has been grouped.
- Discuss your results.

Be creative in your analysis. Consider, for instance, comparing how the popularity of first letters of a name, or the last letters of a name have changed over time. Up to 5 marks will be awarded for each plot, and up to 5 marks for your use of group by or pivot for at least one of those plots.

Bonus marks (optional – up to 10 bonus marks):

Go to the original source in the link above, import the data directly from the URL, and wrangle it in your notebook until your result is the same as the data in 'baby_names.csv'. No marks will be awarded if you have hard-coded values into any cells, if any of the data is pre-wrangled in Excel, or if you do not import the data directly from the URL. We should be able to run your code without reference to any local data files and get the same result as is in 'baby_names.csv'. If you achieve the task, you will be awarded 8/10. If your solution is elegant, you will be awarded 10/10. If you tried but didn't get there, you will be awarded up to 5/10, depending on how well you did, how far you got and how tidy the code you include is.

If you succeed, run these commands and display the output of each one to show that you have achieved the task (assuming your dataframe is called 'Names'):

```
Names.head(10)
Names.tail(10)
Names.year.unique()
Names.sex.unique()
len(Names)
```

Question 3: NZ Indicator data - Cleaning and mini-report (45 marks)

The aim

In this question you will produce a very small piece of analysis. Our aim is that you will be introduced to the idea of producing a structured report that has an overall theme, poses a research question, and does a bit of exploration towards answering that question, interspersed with some very tidy (and hopefully efficient) code. We do not expect that you will exhaustively explore your research question/s.

We also intend that the suggestions below will be a guide to you for when you produce later reports for this course.

(By the way, don't include this aim in your mini-report intro. We will deduct marks if you do. This aim is *our* aim for *you* – the 'meta-aim', if you like. It is not to be confused with *your research goal*.)

The data

The "nz_indicators.csv" dataset is another real-world dataset. It covers socio-economic data on New Zealand, stretching back to early 1980s. The data covers a range of topics: income and wealth distribution, poverty and deprivation levels, health measures, education outcomes, housing as well as employment. The data is captured by various government agencies as well as some private sector entities.

There are approximately ~100 columns in the total dataset. You are being provided with a slightly smaller subset. The columns range widely in their completeness and coverage. A document ("data_documentation.xlsx") is provided which explains briefly what each column means and where it originated. In that document you will see that the columns (or

'features') of the dataset are grouped into the various topics.

The dataset has been intentionally tampered with in order to provide you with practice in data wrangling and cleaning.

The task

For this question, you are to produce a mini-report within your Jupyter Notebook.

Your report should have a structure (as a bare minimum it should have an appropriate title (heading level 1), a short introduction and conclusion, with appropriate formatted headings).

Your report does not have to be in a separate notebook to the answers for your other questions – in fact it is preferred that it is in the same notebook for ease of marking.

To produce the report, your tasks are as follows:

- Look at "data_documentation.xlsx". Pick **two topics** that you would like to investigate and formulate a question or questions for analysis. You may like to investigate, for instance, certain correlations or trends. In your introduction, you should describe what you are looking into, and how you are going to go about it.
- **After your intro, import and thoroughly clean the data** in **all** columns in "nz_indicators.csv" (even the ones you will not be using – we want to see you scrub some data!). You can do this any way you like; however, it is possible to achieve this with a very minimal amount of code so don't be scared by the number of columns. If you are *really* stuck, I have described my cleaning process below. This is only one possible approach, and we love it (and prefer it) when you go your own way. I:

    - Cleaned up nulls on reading in the data
    - Parsed dates in 'year', set is as the index, and then sorted the index
    - Defined a single function that cleans nuisance characters out of a single string, converts the result to a float and returns the result. To manage the different types of data that could pass through the function, I had to use a control flow statement that had an if (of course), one elif and then an else. Hint: my first condition was: *if type(x) == float: return x.*
      (You may find that using regex helps with your cleaning function. You would have to import the 're' module to work with regular expressions.)
    - Passed that function to an applymap() that I used over the entire dataframe at once.

    Use a markdown box to describe up front what cleaning you will be doing before you do the cleaning. You can display a little bit of output to show the progress of your cleaning – but show no more than three or four dataframe snippets (and only show a few rows). Large and/or frequent dumps of data output will be penalised.

    Clean all of the columns, whether or not you plan to produce visualisations from them.

    Don't break your cleaning up into too many code boxes as this disrupts the flow – put it into a maximum of three to five code boxes. After you have inspected data types and nulls or anything else that produces lots of output that is not strictly necessary, it is a good idea to comment that code out and run the box again to suppress the output.

    You should do many different checks to ensure that your cleaning has worked. One such check is counting the amount of nulls in the dataframe before and after cleaning. Sometimes when we clean up strings we can inadvertently return a null value when we should be returning a real value.

    Only include the code that matters. If some code is a dead end for you, bid it farewell/organise a memorial for it/save it somewhere else for later if you think it might be useful, just don't leave it in your report.

- Consider also creating new features that are derived from the original features and using these in your analysis. You don't have to, but it is a cool thing to do.
- Produce four visualisations (plots) that use features from both of your chosen topics and are aimed at answering your question/s. You are welcome to produce more, but only the first four will be marked. To produce your visualisations, it is possible that you will need to wrangle your data a bit more beyond the initial cleaning, depending on the type visualisation you wish to produce.

- In markdown boxes you should have:
    - an **introduction** where you talk a little about the data and introduce your research questions (you absolutely **must not** copy and paste any text from this assignment specification into your report intro)
    - **discussion** of your results – including some discussion after each plot is fairly standard approach, and
    - a brief **conclusion**

  As this is a mini-report, please keep your writing brief.
- Otherwise, how you structure any other part of your report is up to you. Generally such a report will have the code and markdown boxes interspersed. And remember, if you haven't used headings, you don't have structure.
- Finally, due to previous trouble with plagiarism, we will not be providing a model report, even on a different topic.

Marking criteria for question 3:
Marks will be awarded for different components of question 3 using the following rubric. Future assignments will be like question 3 (but more in-depth and will also involve machine learning) and will also follow a marking rubric similar to the one below:

Component	Marks	Requirements and expectations
Data Wrangling	15	Thoroughness in data cleaning, use of a user-defined function. (Five marks will be deducted if no user-defined function is used.)
EDA/Visualisation	16	Appropriate range of visualisations and their effective communication. Marks will be awarded for plots that have appropriate axis labels and titles and encode data appropriately.
Data Analysis	10	Were you curious about the data? Did you ask interesting questions of it? Were the plots appropriate to answer those questions? Did you discuss your results?
Presentation	4	Does your mini report have structure? Does the report flow? Was there appropriate use of markdown? Did you check spelling?

Dataset Usage Conditions:

The dataset is a subset of data that was collated by a group of researchers belonging to the Knowledge Exchange Hub at Massey University. The dataset values are obtained from a mixture of publicly available sources as well as confidential private sources. It also contains several derived values. The dataset represents an important piece of ongoing research and therefore the integrated dataset is the intellectual property of the researchers involved in this project. Thus, the publication or sharing of the dataset beyond this course is not permitted and it must be kept strictly confidential. Failure to adhere to the conditions will result in disciplinary action.

Hand-in: Submit your notebook file via the Stream assignment submission link. Run your notebook before submitting so that all output is visible. Also make a HTML download of your notebook with its output showing and submit that (in case of any issues in opening and running your notebook).