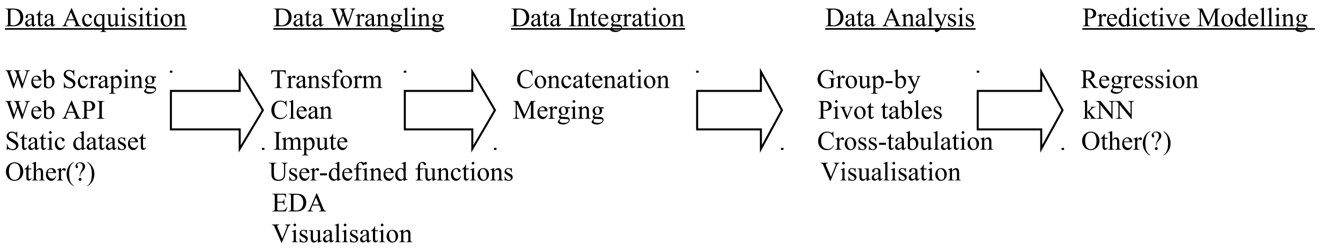
Project 2

Deadline:	Hand in by midnight Sunday, 24th of April 2016.
Evaluation:	15% 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:	Implement the entire data science/analytics workflow. Learn to correctly apply and reason about using different machine learning techniques to solve real-world problems. Gain skills in extracting data from the web using APIs and web scraping. Build on the data wrangling, data visualization and introductory data analysis skills gained up to this point as well as problem formulation and presentation of findings. Learning outcomes 1 - 5 from the course outline.

Project outline:

This project requires that you apply machine learning techniques taught so far to build predictive regression models on current (real-time if possible) data from your chosen domain. You are expected to carry out an entire data science/analytics workflow by: (1) acquiring data from multiple sources, (2) performing data wrangling, (3) integrating the data, (4) conducting analysis to answer some key research questions and finally (5) perform predictive modelling.



Data Acquisition	Data Wrangling	Data Integration	Data Analysis	Predictive Modelling
Web Scraping Web API Static dataset Other(?)	Transform Clean Impute User-defined functions EDA Visualisation	Concatenation Merging	Group-by Pivot tables Cross-tabulation Visualisation	Regression kNN Other(?)

The data should primarily come from up-to-date sources such as web APIs or scraped web pages, but can also be combined with static datasets found in various repositories.

You are predicting continuous valued outputs and are entirely free to choose a domain or a combination of domains. You may for example build predictive models for the housing market in various cities/suburbs, various products, stock prices etc.

Project Requirements:

Project details:

- Each student must work on prediction analysis from different domains/data sources. Therefore, once you have chosen your domain, you must register it on a Google Docs document linked on Stream. Do this as soon as possible in case someone else wants to do the same domain/data source (first-in-first-served).
- The data should be as current as possible and pulled from up-to-date sources.

Suggestions and questions to consider in you experiments:

- Build regression and kNN models and compare their outputs.
- Experiment with models using different feature types. Which features are most effective? Why?
- Experiment with kNN using different distance metrics and different values of *k,* and compare. Which values of *k* are most robust for the size of your dataset and your problem domain? Are variables in your data having different scales affecting the algorithm's accuracy? How have you tried to overcome this?
- Experiment with linear, multiple linear and polynomial regression models and compare. At what point does a regression model become too complex and no longer captures the true relationships in the data?
- How reliable are your prediction models? What do the confidence intervals and prediction bands tell you? Could you recommend this predictive model to a client? Would you expect this model to preserve its accuracy on data beyond the range it was built on?

Submit one IPython Notebook that contains your most integral parts of analysis, together with thorough description of findings. **The Python code in the notebook must be entirely self-contained and all the experiments and the graphs must be replicable**.

Do not use absolute paths, but instead use relative paths if you need to. Consider hiding away some of your Python code in your 'final notebook' by putting them into .py files that you can import and call. This will help the readability of your final notebook by removing unnecessary python code that can clutter and distract from your actual findings and discussions.

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 in order to make your project repeatable on my computer, should I need to install extra modules.

Your notebook must have a heading, abstract (a brief summary of your project together with key findings), an introduction to your research context and research questions, data sources, then the body of your experimental findings, explanations and discussions of the findings and a conclusion. Run your text through an IPython Notebook spell-checker extension.

NOTE: Topics of web scraping, using web APIs and kNN algorithms will be covered in weeks 5 and 6. Therefore, begin your assignment as soon as you can using concepts covered thus far. Once material in weeks 5 and 6 is covered, you will be able to complete all remaining components of this assignment.

Marking criteria:

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

Component	**Marks**	**Requirements and expectations**
Data Acquisition	15	Diversity of sources: a static dataset, data from a web API and data scraped from a web site.
Data Wrangling	15	Quality of your EDA, thoroughness in data cleaning, visualisations.
Data Integration	10	Appropriate use of merging and concatenation.
Data Analysis	20	Quality of the questions being asked, diversity of techniques used to answer and present them.
Predictive Modelling	40	Divesity of experiments. Quality of the evaluation, comparisons and interpretation of results.

Hand-in: **Zip**-up all your **notebooks, python files and dataset(s)** (if there are any static ones) you have chosen into a single file. Submit this file via **stream.**

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