School of Computing and Information Systems
MAST30034: Applied Data Science
Assignment 1
Due date: No later than 11:59pm on Tuesday 13th August 2019
Weight: 20%

Project Overview

The aim of this project is to gain an initial insight into the data set we will be using throughout the subject. This will be achieved through performing an initial analysis, along with a visualisation of the results. The data set we will be using throughout will be the New York City Taxi and Limousine Service Trip Record Data. The data set covers trips taken in various different types of licensed taxi and limousine services in the New York City area. The data is freely available to download from `https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page`. The whole data set is large, covering many years, you are not expect to analyse it all, only a subset that you are free to choose. In this project we want you to pick an attribute to conduct a basic analysis on, and to visualise the results.

You are free to choose the tools you use to perform the analysis and generate the visualisation. You will be required to prepare a report of up to 15 pages detailing the steps taken in performing your analysis and the output of your visualisation.

Project Details

You are free to select a period of time, i.e. month(s), to analyse, as well as the type of licensed taxi you wish to focus on. Your report should explain and justify your selection decisions. Once you have selected your data you should choose an attribute to analyse. You are free to select an attribute that you believe is both of interest, and suitable for visualisation. A simple example would be to analyse the Tip_amount field in the Yellow Taxi data set to determine if different pick-up locations yield different levels of tips. In such an example you would need to first perform a data pre-processing step in

order to extract just data for credit card payments, since only trips that were paid for by a credit card include a Tip_amount. Equivalent pre-processing and cleansing may be required to analyse your chosen attribute.

Once you have performed your analysis you should move to the visualisation stage. You should visualise your analysis onto a map of New York, the type of visualisation will be dependent on the attribute you have chosen, but usage of some form of mapping is required.

The minimum requirement is to produce a geospatial visualisation of a single attribute within the New York City Taxi and Limousine Service Trip Record Data. More marks will be awarded for visualisations that combine multiple attributes, for example, Tip_amount and Trip_distance; with the highest marks available for visualisations that combine additional data sources. For example, evaluating taxi usage around major sporting events or during different weather conditions. Some useful links are provided at the end of this document. Note: when combining multiple datasets the visualisation does not need to be exhaustive, i.e. over multiple months or years, the objective is to determine if there might be a link between the external data and your chosen attribute and to visualise it in such a way as to guide where further analysis could be performed.

Report

Your report should be a maximum of 15 pages and cover at least the following items:

- Data period selection
- Attribute/data selection
- Data pre-processing performed
- Data cleansing performed
- Findings of analysis and description of visualisation

Submission details

Submissions should be made via Turnitin on the LMS.

- Late submissions will incur a deduction of 2 marks per day (or part thereof).

- If you submit late, you MUST email the subject co-ordinator, Chris Culnane
  cculnane@unimelb.edu.au.

Extension policy: If you believe you have a valid reason to require an extension you must contact the subject co-ordinator, Chris Culnane cculnane@unimelb.edu.au at the earliest opportunity, which in most instances should be well before the submission deadline.

Requests for extensions are not automatic and are considered on a case by case basis. You will be required to supply supporting evidence such as a medical certificate. In addition, your git log file should illustrate the progress made on the project up to the date of your request.

Plagiarism policy: You are reminded that all submitted project work in this subject is to be your own individual work. Automated similarity checking software will be used to compare submissions against each other and known public source code. It is University policy that cheating by students in any form is not permitted, and that work submitted for assessment purposes must be the independent work of the student concerned.

Assessment

Your report will be assessed across a number of areas, including:

- Justification of data and attribute selection
- Appropriate pre-processing and cleansing
- Quality and clarity of visualisation
- Analysis of results
- Quality and clarity of report

As already described, the minimum requirement is a geospatial analysis of a single attribute, with more marks available for multiple attribute analysis, and the highest marks available for an analysis that includes some external data.

Useful Links

The data is available from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Further information is available as follows:

- Data Dictionaries
  - Data User Guide: https://www1.nyc.gov/assets/tlc/downloads/pdf/trip_record_user_guide.pdf
  - Yellow Taxi: https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf
  - Green Taxi: https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_green.pdf
  - FHV: https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_fhv.pdf
- Visualisation Tools
  - R: https://cran.r-project.org/doc/contrib/intro-spatial-rl.pdf
  - Python: GeoPlotLib: https://arxiv.org/pdf/1608.01933.pdf, basemap: https://jakevdp.github.io/PythonDataScienceHandbook/04.13-geographic-data-with-basemap.html
- External Data Sources (not an exhaustive list)
  - Weather: https://www.wunderground.com/history/daily/us/nj/newark/KEWR/date/2015-7-28
  - Weather: https://www.timeanddate.com/weather/usa/new-york/historic?month=3&year=2014
  - Baseball Fixtures: https://www.baseball-reference.com/teams/NYM/2015-schedule-scores.shtml
  - Past Events: https://www.nycinsiderguide.com/new-york-city-events-may-2015.html