School of Mathematics and Statistics
MAST30034: Applied Data Science
Assignment 1
Due date: No later than 5:00pm on Friday 4th September 2020
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 attributes to analyse. You are free to select attributes 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 additional chosen attributes. Consider

the pre-processing steps you have learnt in your previous subjects if you are unsure on how to get started.

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 ONE attribute within the New York City Taxi and Limousine Service Trip Record Data, and analysis of TWO other attributes (geospatial visualisation not required).

Report

Your report should be a maximum of 15 pages and cover at least the following items be in either PDF or HTML format. Codes used to provide output should also be submitted.

Submission details

Submissions should be made via Turnitin on the LMS.

- A penalty of 10% of the available marks will be deducted for each day or part-thereof that the submission is late.

Extension policy: More details on the process of applying for a penalty waiver can be found on `https://ask.unimelb.edu.au/app/answers/detail/a_id/5667/~/applying-for-an-extension`

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 according to the following checklist:

Data and Attribute Selection (4 marks)	☐ Clearly states data period (1m) ☐ Clearly states the three (or more) attributes to be analysed (1m) ☐ Convincing justification for data period (1m) ☐ Convincing justification for three (or more) attributes to be analysed (1m)
Pre-processing and Cleansing (4 marks)	☐ Clearly states pre-processing and/or feature engineering steps (1m) ☐ Clearly states data cleansing steps (1m) ☐ Adequately investigates data for possible anomalies/outliers (1m) ☐ Appropriate justification for pre-processing steps, as well as steps for handling missing data (1m)
Visualisation: Quality/Clarity (6 marks)	No marks possible without geospatial visualisation ☐ Geospatial visualisation is present (i.e heatmap, choropleth) (1m) ☐ Appropriate granularity. Is it easily understandable to see what the visualisation is trying to show? Are there too many data points? (1m) ☐ Geospatial visualisation clearly expresses a story, particularly if it raises "interesting" areas of further analysis or indicates an area that does not need further analysis (1m) ☐ Appropriate choice of dimension, colour scheme, legend and formatting (1m) ☐ Appropriate explanation of what the visualisation shows without being overly verbose (2m)
Analysis of result(s) (3 marks)	☐ An appropriate summary statistic to describe the chosen attributes (1m) ☐ Appropriate analysis of the relationship between the attributes (2m)
Quality and clarity of report (3 marks)	☐ Quality writing, spell-checked, correct grammar, and comprehensible sentence structures (1m) ☐ Identifies potential stakeholders, motivation for the report and real-life use cases (1m) ☐ Provides recommendations for potential stakeholders based on analysis of findings (1m)

As already described, the minimum requirement is a geospatial analysis of a single attribute (maximum number of marks is reduced), 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-wi
    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