Assignment 2: Spatio-temporal Data - Group Submission - Returned

Title	Assignment 2: Spatio-temporal Data
Groups	Group B"
Students	Nikita Kapoor (nikitakapoor), Mona Malik (monamalik27), Suseela Pattamatta (suseelapattamatta), Sannath Vemula (sannathreddyv)
Submitted Date	Oct 20, 2019 11:43 PM
Grade	**3.30 (max 4.00)**

Instructions

Objective

Let's play Data Science again! You are provided some spatio-temporal data and asked to illustrate some exciting insights hidden within it. You have two weeks and one day, but you should demo a prototype visualisation to the whole class during the lecture prior to the due date. (Note that this occurs before the midterm *and* the Thanksgiving holiday.) There is no pre-defined "correct solution", but you need to demonstrate mastery of spatio-temporal data modelling in the context of doing Data Science.

Data

The New York City Taxi dataset (https://data.cityofnewyork.us/Transportation/2018-Yellow-Taxi-Trip-Data/t29m-gskq) is a publicly-released dataset of taxi trips in New York City, with a strong spatio-temporal aspect. The full dataset contains over 112 million rows and 17 attributes; I have prepared an archived subset of the first 2^{20} rows, which should suffice for this assignment. You can find it attached. You may wish to create smaller subsets for your initial prototypes. Note that the website includes a link to a pdf that describes each of the 17 attributes. You will also need the taxi zone maps, provided at http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml, in order to interpret the spatial attributes.

Feel free to *integrate* with other map data, such as OpenStreetMap, for alternative visualisation possibilities.

Be certain that you visualise spatial, temporal, or spatio-temporal aspects of the data. Generic multi-dimensional data does not alone fulfill the expectations of the assignment.

Implementation

It is expected that you will use Python, as it is the language of choice in Data Science. Moreover, it provides many visualisation libraries that will help you achieve more than you realise you can! Also, by now you should be more comfortable with it.

Pre-submission Demo

In class on 7 Oct, each group will have 3-5 minutes to demo their prototype to the entire class, using a different presenter than the previous assignment. It is noted that you have little time, but still try to focus the demo on something that you think will generate rich discussion. You may find it helpful in this early stage to leverage the built-in Tableau widget on the dataset webpage. Please send a pdf with at most 4 pages by email to me prior to class so that we do not need to connect laptops to the projector. Also, a discussion will be taken collectively *after the last group presents*; so, a more exciting demo will likely garner more peer feedback.

Submission

You should submit your implementation in raw python or python notebook and a short technical report (5-10 pages) in pdf format. (Note that the increase in page limit is to reflect the submissions for Assignment 1, not an increase in expectation.) The markers will primarily grade the submission based on the report, but may validate that the implementation works as described. The report will describe:

- the insights revealed by the visualisation/application
- the design choices in the implementation/visualisation
- challenges encountered (if relevant)
- how this submission meets the requirements set out in the rubric (i.e., a justified self-evaluation)
- other details that you consider relevant

Assignment 2 Rubric

Component	Weight	4	3	2	1
Transforming Raw Data to Insights	40 %	Visualisation is very informative and tells a story about the data; visualisations are clear and easy to interpret without aide.	Visualisation is informative, albeit possibly with some support from text	Visualisation is complete; information is presented, but lacks complexity or depth.	Very difficult to extract any insights from the visualisation, or no visualisation at all
Data Modelling	30 %	Demonstrates a clear mastery of modelling spatio-temporal data; modelling of the data is flawless and/or demonstrates knowledge of both the strengths and limitations of the data model.	Schema is appropriate for the underlying dataset(s) and successfully incorporates taxi zone map.	Spatial or temporal data model exists, but is not used to produce final visualisation	Minimal use of concepts in spatio-temporal data modelling
Algorithmic Considerations	20 %	Compelling evidence is provided to show the scalability of the submission, by design; visualisations are constructed from the full 2^{20} sample.	Algorithmic considerations are well reasoned and demonstrate that performance has been a design consideration.	Some attempt is made to address questions of efficiency, but the visualisation is restricted to small datasets	Minimal attempt to design an efficient solution; visualisation perhaps does not load

Relationship to Spatio-temporal Challenges	10 %	Research and/or industry challenges for spatio-temporal data model are clearly addressed, adding substantial value to the project.	Research and/or industry challenges for spatio-temporal data model are considered, adding value to the project.	Some attempt is made to engage with research papers presented in class, but it adds limited value to the project	Does not engage with research papers presented in class nor other typical challenges faced with spatial/temporal data

Grading
The entire group will receive a grade based on how well the submission adheres to the rubric below. Note that your report provides the opportunity to persuade the markers, but that they will ultimately grade according to the rubric.

The demo on 7 Oct will *not* form part of the grade; however, it is an opportunity to solicit peer and instructor feedback a few days before the deadline that may ultimately improve the quality of your submission (i.e., your grade). Also, having content to present will contribute to your participation grade in the course.

Tips
1) Try to load the data into a usable structure early (first few days). You may find it more difficult than you expect to model the raw datasets as workable data structures.

2) Ascend quickly to a working visualisation and add complexity later; i.e., first build a minimal viable product (MVP). You will be exposed to new research and new ideas in class as you work on the assignment, so you want to be *agile* with your development patterns. At the same time, even in groups, it may take longer than you expect to go from raw data to a working visualisation; so, you don't want to leave this until the last few days before the demo or you could end up with nothing to show at all!

3) Use libraries prolifically.

4) Consider addressing each rubric component individually in your tech report.

5) Reflect closely on the feedback for Assignment 1. This is meant to guide you towards a higher grade on subsequent assignments.

Extra Resources
Over the course of the next two weeks, supplemental resources (e.g., instructions for Python libraries) *may be* added to the Resources/ panel of the course connex page.

Additional resources for assignment

- taxi-sample.zip (14 MB; Oct 2, 2019 5:20 pm)

Submitted Attachments

- Code and Graphs.zip (4 MB; Oct 20, 2019 11:43 pm)
- CSC 501 Assignment-2 report .pdf (993 KB; Oct 20, 2019 11:43 pm)

Additional instructor's comments about your submission

Overall: 4/3/3/2 = 3.5

Derivation of Insights
+ Visualisations are striking and informative

Data Modelling
+ Very solid (star schema) data model that incorporates spatial, polygonal objects as one fact table and temporal objects as the other
- Insufficient discussion of the spatial/temporal components of the data model, in terms of decisions and trade offs, to demonstrate *clear mastery*; discussion mostly focused on relational component of the model.

Algorithmic Considerations
+ Performance is compellingly shown to be fast, with reported running times on each visualisation under 1s.
= I don't remember teaching data chunking in this class.
- Would be helpful to know how the performance varies as a function of the dataset size. Would this be <2s on twice as much data or <1/2s on half as much data?

Connection to Research
+ Uses Microsoft PowerBI, the tool described in one of the research papers, to derive additional insights
= Some informative discussion of challenges encountered in the project, but these are tool-specific
- Engages well with the tool studied in the Microsoft paper, but not with the ideas presented in the paper (e.g., *what is* a QuickInsight?)