First year project 1, Spring 2021

Road collisions analysis, Michael Szell

Overview: Urban / Spatial Data Science

In this project, you will complete tasks similar to data scientists working for a department of transport or a city government, to inform city leaders about traffic fatalities and injuries, and give insights for urban transport planning. You will explore the latest data set of all recorded road collisions in Great Britain in the year 2019 provided by the UK Department of Transport with details about the circumstances of personal injury.

The major parts of the project are:

- Exploring and transforming the data, making numerical and visual reports
- Connecting data tables (accidents, vehicles, casualties)
- Investigating possible statistical associations by filtering for a variety of attributes
- Visualizing the data on a map
- Involving self-obtained external data sets in the analysis

Requirements

In addition to the requirements in the course description, you must work on github. All Python packages are allowed, but comment your code assuming the reader has no knowledge of extra packages that were not covered in the first semester, such as pandas.

Assignment

You are the data scientist team working for the city hall of a major city in the UK:

City	Birmingham	Leeds	Sheffield	Bradford	Liverpool	Manchester	Bristol
Group	1, 8, 15	2, 9, 16	3, 10, 17	4, 11, 18	5, 12, 19	6, 13, 20	7, 14, 21

Your job is to create a report to the city's lead planner, its lead traffic engineer, and the mayor, informing them about the state of road safety in the city.

Task 0: Data filtering and cleaning

Out of the raw UK data, create a processed data set that is restricted to your city. You are going to use this processed data set for all the tasks below. This data set should contain only fields and records that are relevant to your analysis. Briefly describe your data set in a numerical summary (e.g. number/meaning of tables, fields and records, statistical key metrics).

Task 1: Single variable analysis

Report the frequency of road collisions in your city for 1) different age groups, 2) in different times (during the day, the week, or year), 3) and for differences in one other condition.

Task 2: Associations

Research whether there is a significant statistical association in your city, either 1) between some vehicle attributes and accident circumstances, or 2) between some casualty attributes and accident circumstances. Report whether there is a statistically significant association between such variables or not, together with the appropriate statistical metric(s). Discuss why this association, or the lack of this association, is relevant for urban/transport planning.

Task 3: Map visualization

Visualize the reported collisions on a map of your city. Make a visual distinction between different classes of casualty severity.

Task 4: Open question

Use the data to formulate, motivate, answer, and discuss another research question of your choice. For example, compare your city to the whole UK, or investigate different collision participants (bicycle vs car, bicycle-motored two-wheeler, pedestrian vs car, car vs car,…), or compare the age distribution of casualties to the existing age distribution, or rank/compare how problematic different driver demographics are (young/old male/female),…

For the Tasks 1,2,3,4 you must provide exactly one figure in your report, apart from a textual description, so the report will contain exactly 4 figures (figures may have subplots).

Hand-in

You must hand in:

- gitlog.txt: Your repo's git log, e.g. by running: `git log > gitlog.txt`
- code.zip: One zip file containing one Jupyter notebook (.ipynb) of your commented code that runs fully without errors using the three raw data files, and reproduces your findings. Do not include the three raw data files here. If your code is making use of external data sets or .py scripts, include them here.
- report.pdf: A project report:

The project report must be between 3 and 5 pages long including figures (with 11 pt font size and about 1.5 cm margins, like in this document), and should consist of precisely the following sections:

1. Introduction: Here you provide the context and motivation for the problem. What are your research questions, and why does your research provide value to the city?
2. Data: Here you describe all your data tables/sets, and briefly summarize how you obtained and cleaned/transformed them without referring to any code. You should also describe here how you dealt with data issues such as missing data.
3. Results and discussion: Here you provide the technical results and a discussion of your findings over the data.
4. Limitations: Here you give an account on the major short-coming(s) of your methodology / data.
5. Concluding remarks and future work: Here you provide a couple of sentences summarizing the results of the project, why your report is relevant for your city's urban/transport planning, and indicate how the methods, data, and analysis could be improved or extended.
6. Disclosure statement (optional): Here you may state if there were any serious unequal workloads among group members.

Your whole report will contain exactly 4 figures, and can have max. one table with numerical results. In the report cite at least one literature reference[1], for example in the introduction and/or conclusions.

Your hand-in should be self-contained. You are required to master your code for the oral exam. Your 10 minute oral presentation should correspond to the structure of your report. However, you are encouraged to have slide headings that are more communicative.

[1] You could do it in a footnote like this, or using LaTeX. A citation would look like this, for example:
ITF (2020), *Monitoring Progress in Urban Road Safety*, International Transport Forum Policy Papers, No. 79, OECD Publishing, Paris.