SCHOOL OF INFORMATION AND PHYSICAL SCIENCES

INFT6201 - BIG DATA ASSESSMENT 3: DATA ANALYSIS REPORT

OVERVIEW

- Weighting: 30%
- Due date: Week 12 – 11:59pm, Sunday 13 Nov 2022
- Method of submission: group assessment submitted via Turnitin in Canvas
- Content: Data analysis report
- Length of submission: 12–16 pages (+references and appendix)

DESCRIPTION

Student groups analyse a real dataset by using a professional statistical data analysis tool. The submission includes written text, figures, and tables. References and statistical results need to be formatted in APA style (http://apastyle.apa.org). The Python-code is to be submitted as an appendix (Appendix A). Appendix B is optional and can be used to provide the reader with supplementary material (e.g., figures, tables, supporting documentation). The data analysis report includes:

- Title page
- Executive summary
- Introduction
- Dataset Description
- Descriptives
- Analysis & Results
- Discussion & Conclusions
- References
- Appendix A – Python-code
- Appendix B – Supplementary Material (optional)

To submit your data analysis report, please log on to Canvas and look up the following folder: **Assignments → Assessment 3: Data Analysis Report**

TOPIC

In this assignment, you will be analysing a dataset containing information for traffic accidents that occurred in the state of New York in 2019 and 2020.[1] There are many interesting aspects in the dataset and you will find a wealth of relationships between different variables that can be analysed. In your report, first give an introduction to the topic (Section: Introduction) and then continue with describing the dataset (Section: Dataset Description). The third section (Section: Descriptives) should then provide the reader with an overview of the different variables and relationships between them. This includes summary tables, correlation tables, and figures illustrating data distribution. Then, select a couple of relationships between different variables in the dataset that you find most interesting and investigate these relationships using the statistical techniques we discussed in the course (e.g., *t*-Test, ANOVA, linear regression). Finally, provide a discussion of the results and a conclusion for the data analysis report (Section: Discussion & Conclusions).

[1] The data was sourced from: https://www.kaggle.com/sobhanmoosavi/us-accidents. Please note that the original data file contains information for 49 states of the USA. However, in this assignment, only the accidents in the state of New York are of interest that occurred in 2019 or 2020.

MARKING CRITERIA

Criterion	Description
1. Report format [5%]	A data analysis report has to be clearly structured and well organised. It has to be designed in a pleasant and appealing way (e.g., by using an eye catcher on the title page, including page numbers, number sections, figures and tables, and by adequately highlighting essential information as well as using footnotes and appendices to add supplementary information).
2. Executive summary [5%]	An executive summary provides a concise summary of the most essential information of the data analysis report. This includes in particular the key findings of the statistical analysis as well as any other important discoveries. It is the first thing a reader will see after the title page.
3. Introduction [10%]	A data analysis should provide a concise introduction into the topic (e.g., half a page) and provide an overview of the analysis that are conducted in the report.
4. Dataset Description [5%]	Provide a concise description of the dataset, including the source of the dataset, the number of observations, the number and types of variables, and missing values.
5. Descriptives (e.g., summary tables, correlation tables, figures) [15%]	The data analysis report includes a section that provides an overview of the data. This includes summary tables for key variables (e.g., mean, median, quartiles, standard deviation, etc.), correlation tables, as well as figures (e.g., correlation plots, box plots and violin plots) to provide an understanding of the data distribution.
6. Analysis & Results (e.g., regressions) [25%]	The data analysis report has to provide a well-structured analysis & results section, which presents the analysis that are conducted (e.g., t-tests, ANOVAs, Tukey-Tests, linear regression, etc.) and the results. This includes written text, tables, and figures.
7. Discussion & conclusions [10%]	A data analysis report has to provide a general discussion of the results presented in the earlier sections of the report. Moreover, it has to provide clear conclusions at the end of the report.
8. Appendix A (Python-Code) [10%]	The Python-code used to present the descriptives and to conduct the statistical analysis is to be submitted as an appendix (Appendix A). The Python-code needs a clear structure and has to be free of errors. Use # to comment what analysis is conducted and to structure the code. Marks will be deducted for poor documentation, structure, and programming style.
9. Writing and referencing [15%]	A major aspect of data analysis is the ability to communicate results and findings clearly and accurately. This criterion includes correctness of grammar and spelling, and appropriate use of references, direct quotes, paraphrasing, etc. However, students should note that there is a fine line between poor referencing and plagiarism, and reports that appear to be plagiarised will be referred to the Student Academic Conduct Officer, with possible outcomes such as a mark of zero for the entire report. Students are strongly advised to repeat the University's Academic Integrity Module, and to be sure never to take text or ideas from anywhere without clearly noting the source.