Project Big Data – Assignment 4 (Report & Presentation)

Deadline for report: Monday, July 1 at 23:59

Assignment Goal:

The goal of this assignment is to provide you with practice and experience in exploring a real-life dataset using Python. As part of this assignment, eight datasets on different topics are provided on Canvas (some datasets contain a single .csv file and others contain a set of multiple .csv files).

The topics covered are:
- European hotel ratings and reviews from an online booking platform.
- Residential property descriptions and prices from Washington DC.
- Flight departures and arrivals from three airports in New York.
- Kickstarter project information and funding received.
- PGATOUR (golf) player performance data.
- Airbnb listings and reviews for the city of Seattle.
- Information about Shopify app store apps and reviews.
- Wine descriptions and ratings from experts.

Your group should choose one dataset and perform a set of analyses to extract knowledge and insights from the dataset that can help us learn about the topics represented in the dataset. Details about the datasets are provided in a PDF document within the package of each dataset. For this assignment, we expect that you will ask interesting questions about the data, perform the necessary data processing, visualize trends and relationships, look for correlations between factors, and perform statistical tests of hypotheses.

Assignment Details:

As specified in the course syllabus, this assignment consists of a presentation (20% of your course grade) and a report (35% of your course grade). The presentation and report are graded separately, and will be based on different (but sometimes overlapping) criteria.

In general, we do not specify which analyses you must perform, but expect that the questions you ask are interesting, the analyses you perform are methodologically valid, and the insights you discover are non-trivial. The datasets provided are rich is questions to investigate, and you are free to ask for feedback from the lecturers or TAs if you are not sure about what to investigate. We suggest that you document all questions that you ask/answer, and all analyses that you perform (assuming they are valid). Even if the tests do not favorably answer your question, or if there is no correlation between some factors, your findings may still be insightful.

You are required to hand in your Python code to show that all data processing, visualizations and analyses have been done in Python. Your Python code will not be graded, but will be checked to verify the claims in your report. Your final report should not contain any python code, but should contain relevant graphs and/or tables for visualization. The structure, content and grading criteria are explained further in this document. The report should be written in English.

The report and your code must be handed in via Canvas before July 1 at 23:59. There will be two separate assignments on Canvas, one for your report and one for your Python code. You must submit one file to each assignment (one python file and one PDF document, respectively).

The report should be formatted as a PDF document and must contain page numbers in the bottom right corner. It should be at least 6 pages and no more than 15 pages (including all text and figures). You will be graded on the quality of your report, and therefore longer reports do not earn better grades. A table of contents is not necessary and should be left out. The report must have a front page containing the names and student numbers of the authors, the title of the report, the name of the course, and the date of submission. Each group should hand in only one solution. If two Python files or two PDF files are submitted, only the ***last*** submission will be graded.

We suggest that you start looking into the datasets and thinking about what questions you want to investigate starting from the first week of this course. Time is a serious limitation and you should be able to start on this assignment shortly after the first lecture.

The presentation will be performed by the students in their groups to the lecturers. Each group is alloted 10 minutes for their presentation and 5 minutes for questions. There will be 4-5 groups per presentation session, during which each group should remain present (to serve as audiences for other groups in their session).

Grading Criteria for Final Report:

1. Quality of written text (10%)
• The text is free from grammatical errors and spelling mistakes
• The text has a clear and logical structure

2. Quality of questions investigated (20%)
• The questions that you investigated are potentially valuable and useful for parties relevant to the topic presented in the data
• There are a sufficient number of interesting questions investigated
• The insights that you identified are supported by the data

3. Quality of data analyses (20%)
• Python code is correct (your code does what you intend it to do)
• The appropriate variables are considered for the intended analysis
• Assumptions for statistical tests have been checked
• The appropriate statistical tests have been applied. When there is doubt, the choice is motivated.

4. Quality of the visualizations (20%)
• All graphs have descriptive labels on their axes
• The values on the axes have units
• The intervals of the values on the axes are suitable
• The graph is clear and legible, and uses an appropriate font size
• When appropriate, graphs contain a legible and clear legend
• The use of color in the graphs is helpful for understanding the graphs

5. Quality of the interpretations of the analyses (20%)
• Interpretations are appropriately nuanced
• Interpretations are adequately motivated
• Alternative interpretations are discussed

6. Formal requirements (10%)
• The report should be handed in as a PDF document
• Both the Python file and the final report have the correct filenames
• The report contains a cover page that includes the names and student numbers of the authors, the

name of the course, and the date on which the report was handed in
• The report contains page numbers in the bottom right corner
• The report is at least 6 pages and no more than 15 pages (including all text and figures)

Grading Criteria for Presentation:

1. Quality of the content (60%)
• The presentation contains clear questions and results.
• The presentation is focused and to the point. It is tailored towards its target audience (students in Business Analytics before they take Project Big Data).
• The motivation for the investigations are clearly presented.
• The presentation contains visualizations of the data that support the main narrative of the presentation. These visualizations should be made with the presentation (projector, etc.) in mind.
• The presentation has a clear and transparent structure.
• The presentation offers relevant points for discussion on the basis of the performed analyses.

2. Presentation skills (30%)
• The speaker speaks clearly, audibly, and with good pace.
• The speaker keeps everyone in the audience engaged (eye contact, etc.).
• The speaker uses his or her hands for non-verbal communication.
• The speaker uses body language to convey confidence.
• The speaker stays within the allotted time.
• The speaker responds to questions adequately.

3. Formal requirements (10%)
• The presentation is accompanied with a slide deck. The first slide contains the title of the presentation, the speakers' names, the date, and the title of the course.
• The slides contain sources where appropriate (e.g., citations, borrowed figures, etc.)