Project Big Data

Assignment 4 (Report & Presentation)

You will continue working with data from the "HUE bedtime procrastination study" for which you will need to write a report (35% of your final grade) and give a presentation (20% of your final grade). A cleaned version of the hue data is available on Canvas (`hue_week_4_2017.csv`), as well as another file that contains data from the poststudy questionnaire that participants filled out at the end of the study (`hue_questionnaire.csv`). This file contains the following information:

`gender`	1 = male, 2 = female
`age`	
`chronotype`	Single item (7-point scale), do you consider yourself more of a morning (1) or an evening person? (7)
`bp_scale`	Dutch version of the Bedtime Procrastination Scale [Kroese et al., 2014]
`motivation`	Questions pertaining to personality traits related to procrastination. Single item (7-point scale), how motivated were you to go to bed on time each night? (1 = not motivated, 7 = very motivated)
`daytime_sleepiness`	Dutch translation of the Epworth Sleepiness Scale (4-point scale from 0-3; 8 questions, values summed) ESS; Johns, 1991.
`self_reported_effectiveness`	Single item (7-point scale), do you feel more rested since the intervention

For the final assignment, you will use Python to examine this postquestionnaire data in relation to the HUE data file, visualize trends and relationships, look for correlations between factors, test for significant differences between groups and build a regression model to predict bedtime delay. You are required to hand in your Python code to show that all transformations, visualizations and analyses have been done in Python. Your Python code will not be graded, however. You will bundle your findings in a neatly formatted final report of at most 6 pages (cover page excluded). This report should not contain any python code. The structure, content and grading criteria are explained further in this document. The report should be written in English. In order to perform the analyses, a number of transformations on the data still need to be done. To help you along, a Python template will be made available with a recommended structure for your Python code. The following steps must be implemented.

- Read the hue data file and the questionnaire data file into two separate pandas DataFrames.
- Create a new DataFrame that contains the following Series:

`ID`	Participant ID
`group`	Participant group (1 for experimental, 0 for control)
`delay_nights`	The number of nights a participant delayed their bedtime (range: 0-12)
`delay_time`	The mean time in seconds a participant delayed their bedtime (total delay in seconds, divided by the number of observations measured for each individual, rounded to nearest second).
`sleep_time`	The mean bedtime in seconds of a participant.

  Set the index of this new DataFrame to 'ID'. Note that there should only be a single row per participant ID.
- Fill this new DataFrame by transforming data from the DataFrame about participants' bedtimes (from the hue data file).
- Merge this new DataFrame with the post-questionnaire data and store the resulting DataFrame in a new variable. Perform this joining operation of the two DataFrames in such a way that the resulting DataFrame only contains IDs that were present in both datasets.
- Use the scipy.stats package to calculate correlations between the following sets of determinants:
  - bedtime procrastination scale (`bp_scale`, a personality trait) and mean time spent delaying bedtime. Use the "Pearson correlation tests" to calculate the correlation.

  - age and mean time spent delaying bedtime. Use the "Kendall rank correlation test" to calculate the correlation.
  - mean time spent delaying bedtime and daytime sleepiness. Use the "Pearson correlation test" to calculate the correlation.

- Use the scipy.stats package to determine whether there are significant differences between the experimental group and the control group in terms of:
  - the number of nights participants delayed their bedtime
  - the time participants spent in bed each night
  - the mean time participants spent delaying their bedtime

  Use knowledge gained in the course 'Statistics' to determine which statistical test is appropriate: the t-test or the Wilcoxon rank-sum test. Explain your choice in the Data analyses section of your report.

- For bonus points: When interpreting the findings from the comparisons, adjust the significance level (alpha) to account for an inflation of Type I errors due to performing multiple comparisons. For example, you could use a Bonferroni correction (feel free to consult a reference work on how to adjust the significance level to counteract the inflation of Type I errors).

- Formulate and concisely argue for a hypothesis about which factor or factors (max. three) you believe would best predict delay_time. Write your hypothesis down in the Data analyses section of your report. Note that you should theorize about why you think these factors might be good predictors, and not use stepwise regression to identify the strongest predictors.

- Use `statsmodels.api` to build a regression model that uses your three hypothesized determinants to predict `delay_time` (see page 1 of this document). Make sure that `delay_time` is not included in the list of predictors. Again, do not use stepwise regression, but use the regression model to test your hypothesis.

- Create three distinct, meaningful, well-crafted visualizations that either provide insight into the data, or help support your conclusions. This means creating three different kinds of plots (not three boxplots, or three scatterplots for example).

- Interpret and discuss your findings in the Discussion section of your report.

- Write a succinct conclusion.

Report your findings in a separate document. This document should contain the following sections:

Introduction (100 words) Describe concisely what your report is about. State the research question or research questions with which the report is concerned.

Data description and exploration (500 words) Describe the variables in your final, merged DataFrame: what does each value represent?

For each variable, report appropriate descriptives (e.g., mean and standard deviation, or frequencies, etc.) Mention strange values, and possible outliers.

This is a good place to include one or more visualizations that help the reader to better understand the data.

Data analyses (400 words) Describe the statistical analyses that you performed, and explain why those were the appropriate tests for the data. Report your findings, but don't discuss them yet. This is another place where visualizations can provide insight into the findings.

Discussion (400 words) Discuss your findings. Are correlations high or low? Does that make sense? Are differences significant? How robust do you think your findings are?

Conclusion (100 words) What conclusions could be drawn from your analyses? Prefer nuanced statements to bold, sweeping claims.

The final assignment should be done in the same groups as in the previous weeks, and must be handed in via Canvas on July 1 by 23:59 hours. There will be two separate assignments on Canvas, one for your Python code and one for your report. 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 have the structure as described above. The word count per section is the maximum number of words allowed, excluding tables, captions and figures (note that this word count is strict). A table of contents is not necessary and should be left out. The report contains page numbers on each page in the bottom right corner. 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 first submission will be graded.

Criteria for final report and presentation

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 data analyses (40%)
    - Python code is complete and correct (see individual assignments for the credits that can be earned per component)
    - The reported descriptive statistics are complete (for example, not just the mean, but also the standard deviation)
    - Assumptions for statistical tests have been checked
    - The appropriate statistical tests have been applied. When there is doubt, the choice is motivated.
3. 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
4. Quality of the interpretations of the analyses (20%)
    - Interpretations are appropriately nuanced
    - Interpretations are adequately motivated
    - Alternative interpretations are discussed
5. 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

Criteria for the presentation

1. Quality of the content (60%)
    - The presentation contains a clear research statement.
    - The presentation is focused and to the point. It is tailored towards its target audience (students Business Analytics before they take Project Big Data).
    - Through the presentation, it becomes clear what the motivation for the research statement is.
    - 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.)