| | |
|---|---|
| Informatics and Computer Science | **21CSAI01I**<br>**Introduction to Data Science**<br>**2021 -2022** |
| Module Title **Introduction to Data Science** | |
| Module Leader Nahla Barakat | Semester **Two** |
| Assessment Name: **Project 2** | Assessment Weight<br>**35% of the total course mark** |
| | |

**Instructions to students:**

1. This is a group (***2 students***) assignment.

2. Submission: The student must prepare a Jupyter Notebook file that includes for each step: (a) code to carry out the step, (b) output showing the output of the code, and (c) a short description of how the code works. In case of providing plots, the student has to provide a short description (2-3 sentences) of what the intent of the plot is (the student has to think in terms of variation, co-variation, central trend, spread, skew, etc.), a short text description of the plot, and a sentence or two of interpretation of the plot (again the student has to think concerning variation, co-variation, etc.). Finally, each student must provide max. 10 slides PowerPoint presentation to present her/his work in a Storytelling way, emphasizing the main findings of each part. The project package (Jupyter, Presentation, etc.) submission is allowed though the LMS (e-learning) only.

3. Assessment: Assessment will be on the Jupyter Notebook submitted, in addition to scheduled presentations and discussion with the course team members. Note that the presentation and discussion will be compensated with 10% out of the aforementioned 35%.

4. Feedback: Personalized feedback will be given through discussions. However, final feedback will be provided through the LMS (eLearning).

5. Along with the submitted assignment, you need to submit: a fully completed and signed *Coursework submission form* and a *Statement of Academic Honesty Form*. You can only submit your own work. Any student suspected of plagiarism will be subject to the procedures set out in the GAR.

**Note:**

*The necessary details, data sources and initial codes (if any) will be provided upon releasing the project to the students.*

**Submission Date: Week 13**

**Introduction**

After data collection, tiding, and answering the proposed questions of project 1, every team will continue with the same topic selected and the relevant data sets, where ***additional data from a related messy HTML pages will be scraped, and the process of dada cleaning, tiding, will be repeated for scraped data. This will be then followed by integrating the cleaned scraped data with the data set of phase 1.*** *Further, the students should answer the remaining, (new) questions through exploratory data analysis visualization, correlations and hypothesis testing.*

**Requirements:**

Using the data sets, you preprocessed in project 1, and the additional scraped data, perform data integration, and if additional cleaning or tiding is needed after integration, this should be done. You should answer at least 3 new additional questions, regarding trends in the data sets. For example, (changes of specific features over time, place, age-groups, gender- groups, country, region, profit, losses, ratings, etc.). Different visualization techniques should be used, with justification of the chosen methods. In particular, ***the following steps should be executed:***

**1-** Scrape the extra data from a related HTML pages. **[4 Marks]**

**2-** Clean and tidy the Messy scraped data. **[4 Marks]**

**3-** Integrate all relevant datasets (scraped and main), considering the best matching between the two data sets, and discuss any possible data loss as result of integration. **[4 Marks]**

**4-** Use the integrated data set to answer your remaining questions **[6 Marks]**

**5-** Based on the previous steps, formulate at least one hypothesis, and execute hypothesis testing to validate that hypothesis; **[4 Marks]**

Make sure that in each step of the above, you transform and/or manipulate some of your data to get it into a form that's suitable for the next step. In the final step your

data should be in the best form to answer your questions or otherwise achieve your objectives;

**Deliverables**

1- **a Jupyter Notebook** file that includes for each step: (a) code to carry out the step, (b) output showing the output of the code, and (c) a short description of how the code works. In case of providing plots, the student has to provide a short description (2-3 sentences) of what the intent of the plot is (the student has to think in terms of variation, co-variation, central trend, spread, skew, etc.), a short text description of the plot, and a sentence or two of interpretation of the plot (again the student has to think concerning variation, co-variation, etc.).

**2- A report containing:**

a. Questions: as in project proposal (possibly modified based on feedback)

b. Discuss in reasonable details; how you went about your analysis, aided with snapshots of most important processing, figures and results from your Jupyter Notebook. **[6 Marks]**

c. Finally, (and most importantly) discuss the conclusions you draw from the answers of your questions.

3- **A PowerPoint presentation** in the form **of data storytelling** to introduce your two projects, and all your findings, which will be the base of the final discussion.

**[7 Marks]**