Homework: Data Collection & Pre-Processing

ISB AMPBA Class of 2026: Deadline: March 9, 2025, 11:55 PM.

Prof. Lakshminarayana Nittala

The homework consists of seven questions. In addition to being comprehensive and accurate on the technical portion, as a business analytics professional, you are also expected to be very professional in your communication. Since analytics is very technical, great care has to be taken while communicating the work to a non-technical audience. Therefore, the homework will also be graded on clarity in observations made, organization of the document and overall professionalism. The homework is for a total of 300 points. 30 points will be for professionalism. The following specific guidelines will be helpful to make your work look professional.

Specific guidelines

1. Homework submissions should be in the form of a jupyter notebook.
2. Format your homework submission to look structured and systematic. Use numbering, different font sizes etc to indicate sections, subsections etc.
3. Clearly demarcate each question and parts within a question.
4. Include comments in the code cell to indicate to the reader the task performed.
5. For formatting help, you can use the jupyter notebooks distributed for each session as a guide.
6. Grading will be done by running your code cell by cell, evaluating the output, and any observations made.

General Instructions

- The assignment submission form should also be submitted along with the jupyter notebook. Your submission will not be considered without the Assignment Submission form being submitted.
- All submissions have to be individual and will be checked through Turnitin for plagiarism. Coding Scheme for this assignment is '2N - B'. Honor Code violations are investigated accordingly.
- Please remember to include your name and ID at the top of the assignment.
- There are a total of 7 questions. Attempt all questions.
- Upload your submissions to the 'Assignment Submission' folder on LMS.
- Any late submission will attract a penalty, as mentioned in the course outline.
- Do NOT submit .zip files; otherwise, the submission will not be considered.

Question 1 (50 pts)

1. Use the API documentation at https://random-data-api.com/documentation hosted by random-data-api.com to create a dataset with the endpoint "users". The data should capture 5000 users. (The API gives a maximum of 100 random user profiles in one shot. You will have to write a loop to generate 5000 profiles). Format the data into a dataframe. (10 pts)
2. Add a column to the dataframe that shows the current age of the user. (5 pts)
3. Generate 5000 data points using the *age* column as follows:

- Generate a random sample of 10 ages from the *age* column. Compute the mean. Store it in a list.

- Repeat for 5000 times and append the sample means to the list. Hint: You can use list comprehension to achieve this is in a single line of code. (10pts)

4. Plot a histogram of the sample means. How is the distributed? Write your observations. (10pts)
5. Display a table showing the number of users generated for each payment method. What do you observe? (10 pts)
6. How many unique employment titles are represented in the dataset? (5 pts)

Note: In question 1, please save your dataframe as a csv file. In case we need it for evaluation, we will ask for it later. Your observations will be graded from the standpoint of your data. It is not based on the data we will generate when we run your code.

Question 2 (30 pts)

1. Code a web scraper that will visit https://www.isb.edu/en/study-isb/advanced-management-programmes.html and extract the following information about each program. (i) Title (ii) Brief description (iii) Duration (iv) Work Experience. (20 pts)
2. Create a dataframe that has the following columns: Title, Description, Duration (in months), Capstone Project (Yes/No), Work Experience. Display the dataframe in your notebook. (10 pts)

Question 3 (20 pts)

1. Select a recent news/magazine article related to business analytics that interests you and make a wordcloud using the template provided in the session 4 jupyter notebook. Provide the full text of the article in the notebook so that we can evaluate your work. (15 pts)
2. Do the prominent words in the wordcloud correspond to the gist of the article? What do you observe? (5 pts)

Question 4 (35 pts)

Import the *ISB_video_data.parquet* file. This file contains information scraped from the videos on the ISB youtube channel. Use the data to answer the following questions:

1. Parse the data and create the following columns: (i) Title (ii) Views (iii) Months since posting (15 pts)
2. Plot Views as a function of age of the video, i.e., months since posting. What observations can you make from the plot? (10 pts)
3. What other question do you want to answer using this dataset? Pose a question that interests you and answer it using your analysis. (10 pts)

Question 5 (40 pts)

Import the *quotes_sentiment_data.parquet* file. This file has sentiment scores generated for 1000 random quotes. The *scores* column is the sentiment score from the *NLTK VADER* module. The *sentiment* column has score from the *stanza* package.

1. Rename the columns to reflect the method used. (5 pts)
2. Extract the *compound* score from the *VADER* method and put it in a new column. (10 pts)
3. The sentiment score for each quote generated by *stanza* has several numbers, one for each sentence. Create an average score by averaging over all sentences. Create a new column to store the average score. (10 pts)
4. Let us examine if the two methods are consistent in scoring the sentiment. Make a scatterplot of *VADER* compound score against the average *stanza* score. What would you expect if the two methods are consistent? What do you observe? What is your conclusion? (15 pts)

Question 6 (50 pts)

The *online_reviews.csv* file is a collection of 2478 online reviews of an assortment of products sold on an e-commerce platform. For each *ProductID*, there is a *Score* , i.e., rating, and *Test*, i.e., review, given by the customer.

1. Perform sentiment analysis of the reviews using the VADER module from the NLTK library. Create separate columns for the Positive, Neutral, Negative and Compound sentiment scores. (10 pts)
2. Explore if the rating given by the customer is correlated with the compound sentiment score. Plot the average compound score for each rating and display as a bar plot. Write your observations. (10 pts)
3. Make similar plots using the other three sentiment measures (i.e., Positive, Negative and Neutral). Write your observations. (30 pts)

Question 7 (45 pts)

Import the file *office_iot.csv*. We will only use data for the single day, February 5, 2015 for this question.

1. Create the dataframe with data from February 5, 2015. (5 pts)
2. Create a bar plot that shows the average temperature when room is occupied and when it is not occupied. Write your observations from the bar plot. (10 pts)
3. Repeat above with average humidity. (10 pts)
4. With average Light. (10 pts)
5. With average CO2. (10 pts)