INDIVIDUAL ASSIGNMENT FOR DATA ANALYTICS IN BUSINESS

ADDITIONAL INFORMATION

2021-22

The two instructor’s data samples are explained in this document. Potential business questions associated with the data sample to finish the tasks specified in the assignment brief are also introduced. You will find the formatting requirements for the assignment in Appendix I.

If you opt for the data sample and questions introduced in this document you won’t earn the bonus marks (i.e., up to 10 points adjustments) as specified in the assignment brief (Page 2 – Assignment Brief Final Remark). However, the bonus marks will still apply if you can extend the instructor’s samples. For example, you might find additional variables on either Sample 1 or Sample 2 to construct a more extensive data sample.

Please note that you are allowed to use the business questions I introduced in this document for your assignment. However, you are strongly encouraged to use different questions and objectives.

**Instructor’s Sample 1 and Indicative Example**

Instructor’s Sample 1 contains all S&P 1500 companies listed in the US stock market over three years 2016 – 2018. The variables definitions of this database can be found in Appendix II.

Given that the threat of climate change has triggered a wider concern of the company’s corporate social responsibility, we have selected a series of important indicators about the firm’s sustainability initiatives, e.g., the firm’s environmental disclosure scores. The practitioners might be interested in a business question that does disclose more about the environmental data can predict better financial performance?

You might refer to the following bullet points about the data analysis idea for more details about the suggested structure in the assignment brief:

*Introduction*

1. You must select a company and describe its background as specified in the introduction. I will take 3M Co as an example in this section.
2. Identifying the competitor sectors – you might use the 4-digit SIC Code of 3841 as the competitor sector *or* the 2-digit SIC Code of 38 for a larger sample observation *or* even the whole S&P 1500 sample as a competitor set.
3. The following analysis and modelling in your assignment will be based on the competitor set you identified.

*Main Body*

1. As specified in the suggested structure, we will mention the selected variables for examining the business question, e.g., the environmental disclosure scores and return on assets as a proxy of financial performance, and other relevant variables to examine the interaction effect or as control variable such as firm size (i.e., number of employees).
2. Using the descriptive analysis, we will compare 3M’s performance with other competitors. For example, is 3M financial performance above the industrial average.
3. To address the business question, we can build a regression model between the environmental disclosure scores and financial performance based on the data of their competitors. Using the predictive function, you might answer in your report that is it worth disclosing more environmental data to enhance their financial performance.
4. You can use the findings you obtained to benchmark the existing findings in the literature or industrial reports for the discussion part. For example, is your finding consistent with the literature of Qiu et al. (2016) who suggest the firms that make more extensive disclosures will yield net positive economic benefits.

*Conclusion*

1. Summarizing the key actions points for managers in 3M to consider based on your analysis of the competitors.
2. You will acknowledge the limitation of your analysis, for example, the sample size of your research might be limited and overlook the customer perspectives. For future research, you might suggest the managers design a questionnaire (with some example scales and questions) to analyze how customers respond to 3M’s environmental protection actions.

**Instructor’s Sample 2 and Indicative Example**

The instructor’s Sample 2 is the AirBnB datasets we have been working on in this unit. You are allowed to use all the syntax and examples we have instructed. However, you will need to follow the suggested structures and assignment requirements to report the analysis.

It will be a good idea to combine different datasets based on the host ID to integrate a new dataset to build your model (please refer to the Week 11 exercise on how to integrate different data frames). For example, a possible business question is that what are (or clusters of) amenities for an AirBnB listing in a certain area/season to predict better electronic word-of-mouth EWOM (i.e., the sentiment values of the review comments)?

You might refer to the following bullet points about the data analysis idea for more details about the suggested structure in the assignment brief:

*Introduction*

1. You must select a current listing in the AirBnB and describe its current business and offerings in the introduction.
2. Identify the competitor set of the listing you selected to construct a new sample from the original dataset. For example, you might consider a group of London listings having a similar price range *or* a group of listings located close to the one you selected.
3. The following analysis and modelling will be based on the sample of your identified competitor set.

*Main Body*

1. Describe how you obtain the main variables, such as the sentiment scores of review comments (please refer to the exercise of week 2 Exercise 02 – [Analyzing text](https://github.com/data-analytics-in-business/week-01/blob/main/exercises/02-Advanced.ipynb)) and dummy variables for the amenities etc.
2. Conducting descriptive analysis of the competitor set and benchmark the performance of your selected listing to the competitors.
3. Performing clustering analysis based on the amenities and creating cluster label variables to identify the cluster group of an observation (refer to week 9 in lecture demo – [shoe sales clusters.csv](https://github.com/data-analytics-in-business/week-01/blob/main/exercises/02-Advanced.ipynb)) for further analysis.
4. Based on the cluster label variable, you will examine the roles of different clusters you explored from the clustering analysis to predict the EWOM of the listings. More specifically, you can build a regression model to examine the relationship between cluster groups and EWOM.
5. You can use the findings you obtained from this report to benchmark the existing findings in the literature or industrial reports for the discussion part. For example, can your findings extend the implications from Lee et al. (2019) about the recommendation for AirBnB hosts?

*Conclusion*

1. Summarize the key actions points for managers of your selected listing to consider based on your analysis of the competitors.
2. You will acknowledge the limitation of your analysis, for example, the accuracy of sentiment analysis to appropriately measure EWOM. For future research, you might suggest the managers design a questionnaire (with some example scales and questions) to survey the customer’s expectations for different amenities or service attributes.

APPENDIX I – FORMATTING REQUIREMNTS

**<Title>   
Instructions for preparing your assignment for   
the Data Analytics in Business:  
(Times New Roman, bold, 18 pt, single space, no capitalization)**

<ATTENTION: remove all blue instructions in angle brackets!>

<2 lines spacing>

**General instructions (Section heading: Times New Roman, bold, 12 pt, no numbering)**

The paper should be prepared using ISO A4 paper size (210 x 297 mm) using Times New Roman (12 pt, justified) font according to the format described in this template. The first paragraph of each section and subsection should not be indented.

Following paragraphs indented 0.5 cm.

Margins should be as follows: top 3 cm, bottom 2 cm, left and right 3 cm each.

<1 line spacing before the subheading and section heading, no line spacing after the heading and subheading>

*Other details (subheading: Times New Roman, italics, 12 pt, no numbering)*

Pages should be numbered in the page footer, including the first page, as demonstrated in this document.

Bulleted or numbered lists should have a hanging indent of 0.5 cm. Avoid using more than two levels in the lists.

Following paragraphs indented 0.5 cm.

Footnotes should be avoided. The corresponding text should be incorporated in the main text.

Equations should be typed flush with the left-hand margin and numbered consecutively with numbers in brackets on the right. Leave single spacing above and below equations, as shown below.

<1 line spacing before equation>

Equation (1):

*N*

Σ *xi = δ*

*i=1*

<1 line spacing after equation>

Figures and tables must be included in the main text and be individually numbered and captioned. Illustrations should be sharp and clear.

The placement of figures, tables, and illustrations should be aligned as centered within the document.

Captions should be placed **below figures** and **above tables**. These are to be typed in Times New Roman 11, in italics, and centered.

Leave one line before and after tables and figures (see examples below).

<1 line spacing before table>

*Table 1 – Caption (Times New Roman, Italic, 11 pt, centered)*

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

<1 line spacing after table>

<1 line spacing before figure>

|  |
| --- |
|  |

*Figure 1 – Caption (Times New Roman, Italic, 11 pt, centered)*

<1 line spacing after figure>

This is an example of main text, and is the last line of this section.

<1 line spacing>

**Conclusion**

The above text summarizes and descriptively demonstrates the main instructions for writing and preparing the full assignment. All that is missing now are guidelines regarding references.

References to other publications should be in Harvard style, Times New Roman, 10 pt, left aligned, hanging indent 0.5 cm. They should contain full bibliographical details. Journal titles should not be abbreviated. References should be presented as follows. Within the text: Author’s last name followed by a comma and the year of publication, all in round brackets; e.g., (Fox, 1994). When the number of authors is three or more, references in the text should follow this example (Fox et al., 1995).

At the end of the assignment, list all references in alphabetical order, using the formats described below. All authors should be included in the reference list.

**References**

For books: Surname, Initials (Year of Publication), *Title*, Publisher, Place of Publication.

Casson, M. (1979), *Alternatives to the Multinational Enterprise*, Macmillan, London. (Times New Roman, 10 pt, left aligned, hanging indent 0.5 cm).

For chapters in edited books: Surname, Initials (Year), “Title”, in Editor’s Name, Initials (Ed.), *Title*, Publisher, Place, Pages.

Bessley, M. and Wilson, P. (1984), “Public policy and small firms in Britain,” in Levicki, C. (Ed.), *Small Business Theory and Policy*, Croom Helm, London, pp. 111-126.

For articles: Surname, Initials (Year), “Title”, *Journal Name*, Volume, Number, Pages.

Fox, S. (1994), “Empowerment as a catalyst for change: and example from the food industry”, *Supply Chain Management*, Vol. 2, No. 3, pp. 29-33.

Electronic sources should include the URL of the web site at which they may be found along with the date of latest access.

APPENDIX II – VARIABLE DEFINITIONS FOR SAMPLE 1

**Ticker (TICKER)**

Ticker is a specific identifier for a financial instrument that reflects common usage. Tickers are not, however, unique to specific exchanges or specific pricing sources. Tickers may change in conjunction with Corporate Actions.

**ISIN Number (ID\_ISIN)**

A twelve-character number assigned by the local national numbering agency. The International Securities Identification Number (ISIN) consists of a two-letter country code, followed by the nine-character alphanumerical national security identifier, and a check digit.

**Year (YEAR)**

A company’s fiscal year.

**Name (NAME)**

Name of the company or brief description of the instrument. The Name of an instrument may change in conjunction with corporate actions.

**SIC Code (EQY\_SIC\_CODE)**

A 4-digit Standard Industrial Classification code. It is the statistical classification standard underlying all establishment-based Federal economic statistics classified by industry.

**GICS Industry (GICS\_INDUSTRY)**

A numeric code indicating GICS industry classification. GICS (Global Industry Classification Standard) is an industry classification standard developed by MSCI in collaboration with Standard & Poors (S&P). The Global Industry Classification consists of 11 sectors, 24 industry groups, 62 industries, and 132 sub-industries. The GICS classification assigns an industry to each company according to its principal business activity.

**Country or Territory of Domicile (CNTRY\_OF\_DOMICILE)**

Specifies the ISO (International Organization for Standardization) 3166-1 alpha 2 code of the location of the company’s senior management.

**Number of Employees (NUM\_OF\_EMPLOYEES)**

Number of people employed by the company, based on the number of full time equivalent. If unavailable, then the number of full-time employees is used, excluding part time employees.

**Total Assets (BS\_TOT\_ASSET)**

The total of all short and long-term assets as reported on the Balance Sheet.

**R&D Expense (IS\_RD\_EXPEND)**

Total research and development expenditures incurred which includes R&D in profit and loss account and capitalized R&D during the period. In the case where total R&D expenditure is not disclosed, this field may return profit and loss account only.

**R&D Expense Adjusted (IS\_OPEX\_R&D)**

Research and development expenses that a company has incurred in the period adjusted by one-time charges such as restructuring, merger and acquisitions, gain on sale of subsidiaries, and other nonrecurring activities.

**Operating Expenses R&D (IS\_OPERATING\_EXPENSES\_R&D)**

Research and development (R&D) expense attributable to operating expenses.

**Cash and Cash Equivalents (BS\_CASH\_NEAR\_CASH\_ITEM)**

Cash and Near Cash Items: Cash in vaults and deposits in banks. Includes ST investments with maturities of less than 90 days. May include marketable securities and short-term investments with maturities of more than 90 days if not disclosed separately. Excludes restricted cash (Restricted cash is included in Other Current Assets).

**Environmental Disclosure Score (ENVIRON\_DISCLOSURE\_SCORE)**

Proprietary Bloomberg score based on the extent of a company’s environmental disclosure as part of Environmental, Social and Governance (ESG) data. Companies that are not covered by ESG group will have no score and will show N/A. Companies that do not disclose anything will show a value of ‘0’. The score ranges from 0.1 for companies that disclose a minimum amount of ESG data to 100 for those that disclose every data point collected by Bloomberg. A consistent list of topics, data fields, and field weights apply across sectors and regions. This score measures the amount of environmental data a company reports publicly and does not measure the company’s performance on any data point.

**Social Disclosure Score (SOCIAL\_DISCLOSURE\_SCORE)**

Proprietary Bloomberg score based on the extent of a company’s environmental disclosure as part of Environmental, Social and Governance (ESG) data. Companies that are not covered by ESG group will have no score and will show N/A. Companies that do not disclose anything will show a value of ‘0’. The score ranges from 0.1 for companies that disclose a minimum amount of ESG data to 100 for those that disclose every data point collected by Bloomberg. A consistent list of topics, data fields, and field weights apply across sectors and regions. This score measures the amount of social data a company reports publicly and does not measure the company’s performance on any data point.

**Governance Disclosure Score (GOVNCE\_DISCLOSURE\_SCORE)**

Proprietary Bloomberg score based on the extent of a company’s environmental disclosure as part of Environmental, Social and Governance (ESG) data. Companies that are not covered by ESG group will have no score and will show N/A. Companies that do not disclose anything will show a value of ‘0’. The score ranges from 0.1 for companies that disclose a minimum amount of ESG data to 100 for those that disclose every data point collected by Bloomberg. A consistent list of topics, data fields, and field weights apply across sectors and regions. This score measures the amount of governance data a company reports publicly and does not measure the company’s performance on any data point.

**Tobin's Q Ratio (TOBIN\_Q\_RATIO)**

Ratio of the market value of a firm to the replacement cost of the firm’s assets. The Q ratio is useful for the valuation of a company. It is based in the hypothesis that in the long run the market value of a company should roughly equal the cost of replacing the company’s assets. The ratio is computed as follows: (Market Cap + Total Liabilities + Preferred Equity + Minority Interest)/Total Assets

**Return on Assets (RETURN\_ON\_ASSET)**

Indicators of how profitable a company is relative to its total assets, in percentage. Return on assets gives an idea as to how efficient management is at using its assets to generate earnings.

**Return on Common Equity (RETURN\_COM\_EQY)**

Measure of a corporation’s profitability by revealing how much profit a company generates with the money shareholders have invested, in percentage.

**Gross Margin (GROSS\_MARGIN)**

Gross margin represents the percent of total sales revenue that the company retains after incurring the direct costs associated with producing the goods and services sold by a company. Calculated as: (Net Sales – Cost of Goods Sold) \* 100/Net Sales.