|  |  |
| --- | --- |
| . | DS1 Introduction to Data Science  Study Guide Bachelor Applied Artificial Intelligence & Data Management  Academic Year 2021-2022 |

Index

[1 General information 2](#_Toc40102014)

[2 Content of the module 2](#_Toc40102015)

[**2.1** **Module description** 2](#_Toc40102016)

[**2.2** **Relation B AAI&DM Competency Profile** 2](#_Toc40102017)

[**2.3** **Learning objectives** 2](#_Toc40102018)

[3 Teaching and Learning Activities 3](#_Toc40102019)

[**3.1** **Teaching and learning activities** 3](#_Toc40102020)

[**3.2** **Assignments** 3](#_Toc40102021)

[**3.3** **Planning** 3](#_Toc40102022)

[4 Study load 4](#_Toc40102023)

[5 Literature 4](#_Toc40102024)

[6 Assessment 4](#_Toc40102025)

[**6.1** **Assessment methods, type of evaluation and weight** 4](#_Toc40102026)

[**6.2** **Grading: Minimum grade, Compensation, and Duration of Validity** 4](#_Toc40102027)

[**6.3** **Assessment criteria** 5](#_Toc40102028)

[**6.4** **Requirements for handing in assignments** 5](#_Toc40102029)

[**6.5** **Grading** 5](#_Toc40102030)

[**6.6** **Resit** 6](#_Toc40102031)

1. General information

Module name: Introduction to Data Science

Module code: DS1

Semester: Block A

Number of ECTS credits: 4 ECTS credits (or 84 hours)

Module Assessment Weight (out of 100): 26 points

Year: 1

Lecturer(s): Bram Heijligers; Nitin Bhushan

Contact details: [heijligers.b@buas.nl](mailto:heijligers.b@buas.nl); [bhushan.n@buas.nl](mailto:bhushan.n@buas.nl)

1. Content of the module
   1. **Module description**

This module introduces students to the fundamental concepts and techniques for extracting useful knowledge from data representations thereof. These concepts are of three types. First, it will discuss data-analytic thinking, and introduce data science standards that are commonly used in the industry. Second, the module introduces students to concepts in statistics and probability theory that form a basis for modern data science. Lastly, the module will zoom in on data analysis, visualization and reporting.

* 1. **Relation B AAI&DM Competency Profile**

Intended Learning Outcome 4. Researching and Analysis:

1) The student demonstrate a full understanding of different data types, descriptive data, graphical representations of data, statistical inference, basic probability theory, correlation and simple linear regression.

2) The student can apply the CRISP-DM, sigma-notations and linear transformations.

Intended Learning Outcome 5. Conceptualizing

3) The student can transform a business requirement into a data science problem and propose an effective solution.

4) The student is able to formulate an insightful data-driven research question, quantify appropriate real-world phenomena and objects into data and covert this data into meaningful graphical representations thereof.

dule addresses the competency ‘………………’

* 1. **Learning objectives**
* Students demonstrate an understanding of different data types. This module introduces students to tabular data and the different data types normally encountered in tabular data. Students learn to distinguish between continuous and categorical data types. Further, they learn the difference between ordinal and nominal data types.
* Students can effectively describe data using statistics and graphs. Students learn to summarize and describe continuous and categorical data types using summary statistics including measures of location (mean, median, mode) and measures of dispersion (standard deviation, variance, range, IQR), and learn how to compute these measures by hand. Further, students learn the appropriate data visualization techniques applicable for each data type.
* Students demonstrate an understanding of statistical inference and basic probability theory. Students learn that data science often involves samples of data. Next, they understand that samples originate from populations which often have a probability model associated with them. They are introduced to basic concepts in statistical inference (to infer population parameters using sample statistics) using the standard normal distribution. Last, students learn the concept of conditional independence and Bayes rule from probability theory.
* Students demonstrate an understanding of correlation and simple linear regression. Further, they learn how to compute a correlation coefficient by hand using sigma notation.
* Students are familiar with CRISP-DM. They understand the different roles and responsibilities involved.
* Students are familiar with sigma-notation and linear transformations. They demonstrate their understanding by computing descriptive statistics and correlation coefficients by hand.
* Students can translate a business requirement into a data science problem. In sum, the student can identify different forms of tabular data and perform appropriate analyses (data description, data visualization, correlation and/or regression).

1. Teaching and Learning Activities
   1. **Teaching and learning activities**

Instruction methods used in this module are lectures, e-learning, workshops and case studies. Students are expected to self-study online study materials (mainly interactive swirl classes and educational videos) before coming to class. In the lectures the main concepts discussed in the interactive workshops and textbooks will be reviewed, and illustrated using examples and short cases. In order to deepen understanding, and train students in the application of the concepts. In the same lecture, the students will be introduced to the interactive workshop and mock assessment which they can continue on campus or at home. Students are encouraged to continue to work after the lecture as they can receive faster feedback; lecturers will respond to questions and provide support during working hours.

* 1. **Assignments**

Each day students have to prepare classes by completing the previous workshop. They are encouraged to make the mock assessment and are given the opportunity to collect feedback from the lecturers on their work to prepare the for the actual conference poster.

* 1. **Planning**

|  |  |  |  |
| --- | --- | --- | --- |
| **Week** | **Day** | **Preparation** | **Lecture-Workshop Hybrid** |
| 3 | 1 | None | Introduction to DS (seeing the world in data, attributes) |
| 3 | 2 | Completed workshop 1 | Variables (data frames, continuous, nominal, ordinal etc.) |
| 3 | 3 | Completed workshop 2 | Descriptive analyses (mean, sd, range, IQR) & visualisation (boxplots) |
| 4 | 4 | Completed workshop 3 | Introduction to Probability (random variable, distributions) |
| 4 | 5 | Completed workshop 4 | Introduction to stat. inference (sample, pop, hypothesis testing) |
| 4 | 6 | Completed workshop 5 | Analysing Relationships between variables (e.g., compute correlation by hand) |
| 5 | 7 | Completed workshop 6 | Reporting & visualising (Academic Skills) |
| 5 | 8 | Completed workshop 7 | Introduction to regression & ML (supervised, unsup..., the CRISP DM model) |
| 5 | 9 | Completed all workshops | Recap and working on creative brief |

1. Study load

The study load for the module is divided as follows:

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Self-study | 0 |  |
| Interactive Workshops (including preparation) | 45 (8x6) |  |
| Mock Assesment (case study) | 16 (8x2) |  |
| Final Poster/Presentation Preparation | 8 |  |
| TOTAL | 72 |  |

1. Literature

* OnlineStatBook: Online Statistics Education: A Free Resource for Introductory Statistics (onlinestatbook.com)
* Learning Statistics with R, by D. Navarro (2018): https://learningstatisticswithr.com/
* Discovering Statistics using R, A. Fields et al (2012): https://uk.sagepub.com/en-gb/eur/discovering-statistics-using-r/book236067
* Swirl package for Interactive Programming Courses in R: https://swirlstats.com/

1. Assessment
   1. **Assessment methods, type of evaluation and weight**

The assessment of the module will take place through:

|  |  |  |
| --- | --- | --- |
| **Type of assessment** | **Type of evaluation** | **Percentage of the final grade** |
| **Individual assessment** | Conference Poster | 100 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

* 1. **Grading: Minimum grade, Compensation, and Duration of Validity**

**Minimum grade**

In order to pass the module, the students need to complete the creative brief and score sufficiently in the assessment criteria of this module.

* 1. **Assessment criteria**

**Individual Assignments (100% of total grade)**

**Create a Conference Poster based on case studies and examples introduced online and in-class; the focus is on all stages of the data science process (formulating the problem; selecting the data and exploratory data analyses technique(s); interpretation of descriptive data; reporting and visualizing the data) from an elementary level with specific emphasis on data visualisation as a means for understanding the data . Propose next steps: an analysis to run on the data.**

* Students can formulate a data-driven research question.
* Students demonstrate an understanding of different data types.
* Students can effectively describe data using statistics and graphs.
* Students demonstrate an understanding of statistical inference and basic probability theory.
* Students demonstrate an understanding of correlation and simple linear regression.
* Students are familiar with CRISP-DM.
* Students are familiar with sigma-notation and linear transformations.
* Students can translate a business requirement into a data science problem.

|  |  |  |
| --- | --- | --- |
| **Component** | **Description of component** | **ILO** |
| 00 Introduction | What is the problem statement? | 5 |
| 01 EDA Methodology | Describe the data. | 4 |
| 02 Findings | Include relevant visuals. | 5 |
| 03 Discussion | Critically look back at your solution . Do we have a data-driven solution to the problem statement? | 4, 5 |
| 04 Conclusion | Summary and closing statement | 4, 5 |
| 05 References, notes and author information | Sources, references, notes and author information. Live presentation of poster on DataLab 05 | 1,2,3 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | |
| ILO | Missing | Poor | Insufficient | Sufficient | Good | Excellent |
| **5. Conceptualizing** | Not addressed this block in your project work. Your project work evidencing can include your Learning Log, Work-log, GitHub commits and supporting documents you submitted with your project by uploading during hand-in. | "The student:  1) Demonstrates a poor or no understanding of different data types, descriptive data, graphical representations of data, statistical inference, basic probability theory, correlation and linear regression.  2) Does not or poorly applies the CRISP-DM, sigma-notation and linear transformations.    by not using the correct data science concepts and R packages out of his repetoire, with a wrong implementation and wrong interpretation. " | "The student:  1) Demonstrates a lacking or fragmented understanding of different data types, descriptive data, graphical representations of data, statistical inference, basic probability theory, correlation and linear regression.  2) Partly applies the CRISP-DM, sigma-notation and linear transformations.    by partly using the correct data science concepts and R packages out of his repetoire, with a wrong implementation and wrong interpretation. " | "The student:  1) Demonstrates a basic understanding of different data types, descriptive data, graphical representations of data, statistical inference, basic probability theory, correlation and linear regression.  2) Applies the CRISP-DM, sigma-notation and linear transformations.    by using the correct data science concepts and R packages out of his repetoire, with a wrong implementation but right interpretation. " | "The student:  1) Demonstrates an full understanding of different data types, descriptive data, graphical representations of data, statistical inference, basic probability theory, correlation and linear regression.  2) Applies the CRISP-DM, sigma-notation and linear transformations.    by using the correct data science concept out of his repetoire, with a correct implementation and interpretation. " | "The student:  1) Demonstrates an comprehensive understanding of different data types, descriptive data, graphical representations of data, statistical inference, basic probability theory, correlation and linear regression.  2) Effectively applies the CRISP-DM, sigma-notation and linear transformations.    by using the correct data science concept out of his repetoire, with a innovative and effective implementation followed by insightful interpretation. " |
| **4. Research and Analysis** | Not addressed this block in your project work. Your project work evidencing can include your Learning Log, Work-log, GitHub commits and supporting documents you submitted with your project by uploading during hand-in. | "The student:  1) Can not or poorly transform a business requirement into a data science problem and does not propose a suitable solution.  2) Is able to poorly formulate an data-driven research question, quantify appropriate real-world phenomena and objects into data and does not covert this data into meaningful graphical representations.    as demonstrated by an unsuitable statistical analysis approach and by creating unappropriate visualisations in R." | "The student:  1) Can transform a business requirement into a data science problem but does not propose a suitable solution.  2) Is able to formulate an insightful data-driven research question, quantify appropriate real-world phenomena and objects into data but does not covert this data into meaningful graphical representations thereof.    as demonstrated by an unsuitable statistical analysis approach or by creating unappropriate visualisations in R." | "The student:  1) Can transform a business requirement into a data science problem and propose a suitable solution.  2) Is able to formulate an insightful data-driven research question, quantify appropriate real-world phenomena and objects into data and covert this data into meaningful graphical representations thereof.    demonstrated by a suitable statistical analysis approach and by creating appropriate visualisations in R." | "The student:  1) Can transform a business requirement into a data science problem and propose an effective solution.  2) Is able to formulate an insightful data-driven research question, quantify appropriate real-world phenomena and objects into data and covert this data into meaningful graphical representations thereof.    demonstrated by a effective statistical analysis approach and by creating appropriate visualisations in R." | "The student:  1) Can transform a business requirement into a data science problem and propose an effective solution.  2) Is able to formulate an insightful data-driven research question, quantify appropriate real-world phenomena and objects into data and covert this data into innovative graphical representations thereof.    demonstrated by a effective statistical analysis approach and by creating innovative visualisations in R." |

* 1. **Requirements for handing in assignments**

The individual assignments are to be handed in Microsoft Teams or Github no later then 5pm on the day before the session in which we present the posters.

* 1. **Grading**

Grading will take place in accordance with the TER. Grades will only be processed if you have registered in Osiris for the examination of the module.

*Missed opportunities*

If you have registered to do participate in the module did not deliver (a part of) the conference poster, your result will be registered as a missed opportunity (abbreviated as GK).

*Invalid sufficient average grades*

In the case in which the final average grade is sufficient, but one of the parts (for example essay or exam) is insufficient, the letter O will be registered in Osiris to indicate that the results are insufficient.

*Rounding and averaging*

Grades will be entered in Osiris up to one decimal.

If marks consist of more than one decimal, the original mark will be broken off after the first decimal. A few examples:

• 5.49 becomes 5.4

• 6.73 becomes 6.7

*For further regulations on grading, see TER of B AAI&DM.*

* 1. **Resit**

Each student receives two exam opportunities per year. The first opportunity is during or at the end of the module (for assignments). The second opportunity follows later. As a rule, the resits for assignments (both group and individual) also take place in the exam period in which the resit of the exam will take place.

You need to register in Osiris to be able to take part in resits (whether exams, assignments, or reports).