Module
INS302 - Data Science

Due date for submission
(see Wiseflow)

Subject's supervisor
Noha El-Ganainy, Tomas Sandnes

Teacher and e-mail
Vahid Hassani | vahid.hassani@oslomet.no

Learning outcomes

After successfully completing the course the student:

Knowledge

- can define the term Data Science
- has knowledge of various data sources, such as database, social media and sensors
- can describe basic visualization principles

Skills

- can use predictive / advanced analysis tools
- masters the tools for visualizing information and facilitating an end user
- can perform selected techniques in statistics and analysis, such as correlation, regression analysis clustering

General Competence

- understand how the whole process from raw data acquisition to data visualization creates value
- can assess ethical issues surrounding data collection and analysis.

Please address the following questions (cases) in your submission.

This assignment must be solved individually and counts for 100% of the total grade. The weighting of each sub-task is stated as a percentage.

You have been given access to this data set and must first spend some time getting to know it.

The dataset refers to sad story of Titanic sinking in 1912 after hitting an iceberg. The disaster killed over 67% of the passengers and crew.

The dataset is in the attached file "Titanic.zip". Open this file in a tool you are familiar with, for example, you can import the content into Excel or use Python packages such as Pandas.

Task 1 - Understanding the data (15%)

- What kind of data set is this? Provide some context by saying a little about what kind of objects we are talking about, and what kind of attributes they have.
- What kind of data types are present in the data set? Provide concrete examples.

Task 2 - Utility value (15%)

- What can you use this data set for? Name at least 2 different applications, or examples of getting value out of the data set.
- Is the data set ready to be analyzed as it is? Is some data processing necessary before using it? Provide arguments that justify your answer.

Task 3 - Analysis and modeling (40%)

- What methods would you apply to analyze the data, given at least one of the applications you have mentioned above? Consider the potential application of machine learning, statistical methods, visualizations, etc. Justify your choices, using bibliographic references from the course literature.
- Partition the dataset into two smaller sets: one for passengers who died and the other one for those who survived. Consider the age of these groups. Plot the histogram of the age for each group. Calculate the interesting statistical properties of each group (age) such as mean, variance, and higher statistical moments. Can you find any special statistical property that differentiate the two groups. Try to interpret the results from your perspective.
- Carry out the analysis suggested above, clearly explaining each step of the procedure you choose to follow.

Task 4 - Results and evaluation (30%)

- What did you get out of the data? Show concrete numbers, figures and graphs.
- Carry out a critical reflection on the result and the insight you have arrived at. Discuss your assumptions and the range of validity of your predictions in detail.

Assignment specification

1. A report (3500 and 5000 words) that answers all the assignment questions. Remember to include illustrations where appropriate.
2. The report must also contain a separate section that shows an overview of which tools you have used in the assignment.
3. All source code you have written must be included as an attachment to the report.

Assignment criteria*

Grade	Learning Outcome 1: Knowledge	Learning Outcome 2: Skills	Learning Outcome 3: Competence
A Excellent	Excellent and comprehensive understanding of concepts	Demonstrates excellent analytical, technical and writing skills	Outstanding degree of judgment and independent critical thinking
B Very good	Very good understanding of concepts	Demonstrates very good analytical, technical and writing skills	Sound degree of judgment and independent critical thinking
C Good	Good understanding of theory in most important areas	Demonstrates good analytical, technical and writing skills	Reasonable degree of judgment and independent critical thinking
D Satisfactory	Satisfactory understanding of theory, but with significant shortcomings	Demonstrates limited analytical, technical and writing skills	Limited degree of judgment and independent critical thinking
E Sufficient	Meets the minimum understanding of concepts	Demonstrates sufficient analytical, technical and writing skills	Very limited degree of judgment and independent critical thinking
F Fail	Fail to meet the minimum academic criteria.	No demonstration of analytical, technical and writing skills	Absence of judgment and independent critical thinking

*Adapted from The Norwegian Association of Higher Education Institutions (http://www.uhr.no/utdanning/karakterpanel_1)

The assignment is worth 100 % of the grade of the course.