Task details

This assessment will give you practical experience in data visualisation, exploration, and preparation (preprocessing and transformation) for data analytics. This assignment is individual work. Each of you will be working with an individual dataset that you can download from the link below.

**Objectives:**

* This assessment task addresses the following subject learning objectives (SLOs): 2 & 4
* This assessment task contributes to the development of the following Course Intended Learning Outcomes (CILOs): D.1

### **Scenario**

You have just started working as a data miner/analyst in the Analytics Unit of a company. The Head of the Analytics Unit has brought you a dataset [a welcome present ;-))]. The dataset includes two files: a description of the attributes and a table with the actual values of these attributes. The Head of the Analytics Unit has mentioned to you that this is some sort of airline customer satisfaction data that a the client Airline\_FLY32130 has provided for analysis. The Head of the Analytics Unit would like to have a report with some insights about that data, that they could deliver to the client. Your tasks include:

* understanding the specifics of the dataset;
* extracting information about each of the attributes, possible associations between them, and any other specifics of the dataset.

The tasks in the assignment are specified below.

For this dataset, you only have the attribute headings and a brief description, which you can find here (attached here as ATTRIBUTES.xlsx)



Each student is assigned an individual table with the actual values of these attributes. You will find your individual dataset attached below as (DATASET.xlsx). Your dataset is the one with your student ID in the file name.



### **Tasks**

#### **1A. Initial data exploration**

1. Identify the attribute type of each attribute in your dataset. If it's not clear, you may need to justify why you chose the type.
2. Identify the values of the summarising properties for the attributes, including frequency, location and spread (e.g. value ranges of the attributes, frequency of values, distributions, medians, means, variances, percentiles, etc. - the statistics that have been covered in the lectures and materials given). Note that not all of these summary statistics will make sense for all the attribute types, so use your judgement! Where necessary, use proper visualisations for the corresponding statistics.
3. Using KNIME or other tools, explore your dataset and identify any outliers, clusters of similar instances, "interesting" attributes and specific values of those attributes. Note that you may need to 'temporarily' recode attributes to numeric or from numeric to nominal. The report should include the corresponding snapshots from the tools and an explanation of what has been identified there.

Present your findings in the assignment report.

#### **1B. Data preprocessing**

Perform each of the following data preparation tasks (each task applies to the original data) using your choice of tool:

1. Use the following binning techniques to smooth the values of the following two attributes:

* **- Departure Delay in Minutes**
* **- Arrival Delay in Minutes**

For each attribute, you must apply:

* 1. Equi-width binning
  2. Equi-depth binning

In the assignment report, for each of these techniques, you need to illustrate your steps. In your Excel workbook file place the results in separate columns in the corresponding spreadsheet. Use your judgement in choosing the appropriate number of bins - and justify this in the report.

1. Use the following techniques to normalise the following attribute:

* **- Flight Distance**

For this attribute, you must apply:

* 1. min-max normalization to transform the values onto the range [0.0-1.0].
  2. z-score normalization to transform the values.

The assignment report provides an explanation of each of the applied techniques. In your Excel workbook file place the results in separate columns in the corresponding spreadsheet.

1. Discretise the **AGE** attribute into the following categories:
   * Children and Adolescents (0-18 years)
   * Young adults (18-35 years)
   * Adults (35-65 years)
   * Elderlies (65-100 years)

Provide the frequency of each category in your dataset.

Your assignment report should provide an explanation of each of the applied techniques. In your Excel workbook file place the results in a separate column in the corresponding spreadsheet.

1. Binarise the **CLASS** variable [with values "0" or "1"].

Your assignment report should provide an explanation of the applied binarisation technique. In your Excel workbook file place the results in separate columns in the corresponding spreadsheet.

#### **1C. Summary**

At the end of the report include a summary section in which you summarise your findings. The summary **is not** a narrative of what you have done, but a condensed informative section of **what you have found** about the data that you should report to the Head of the Analytics Unit. The summary may include the most important findings (specific characteristics (or values) of some attributes, important information about the distributions, some clusters identified visually that you propose to examine, associations found that should be investigated more rigorously, etc.).

### **Deliverables**

The deliverables are:

* A report, for which the structure should follow the tasks of the assignment, and
* An Excel workbook file with individual spreadsheets for each task (spreadsheets should be labelled according to the task names, for example, "1A"). Each of the results of parts (a) to (d) in task 1B should be presented in a separate sheet (and respectively table in the assignment report).

In the report, include a section (starting with a section title) for each of the tasks in the assignment.

## Submission details

**Length:**The task requires the submission of a report (approx. 2500-3000 words or 20 pages using 11 or 12-point Times or Arial fonts). Please also upload the Excel spreadsheet containing your results